DOCUMENT RESUME

ED 395 000 TM 024 996

AUTHOR	Bunderson, C. Victor; And Others
TITLE	The Four Generations of Computerized Educational Measurement.
INSTITUTION	Educational Testing Service, Princeton, N.J.
REPORT NO	ETS-RR-88-35
PUB DATE	Jun 88
NOTE	154p.; In: Linn, R. L., Ed. "Educational Measurement" (3rd edition). New York: Macmillan, 1988.
PUB TYPE	Reports - Evaluative/Feasibility (142)
EDRS PRICE	MF01/PC07 Plus Postage.
DESCRIPTORS	Academic Achievement; *Adaptive Testing; *Computer Assisted Testing; *Educational Assessment; Educational Innovation; Information Dissemination; Information Processing; *Intelligent Tutoring Systems; *Measurement Techniques; Models; Research Methodology; Scoring; *Test Items

ABSTRACT

Educational measurement is undergoing a revolution due to the rapid dissemination of information-processing technology. The recent growth in computing resources and their widespread dissemination in daily life have brought about irreversible changes in educational measurement. Recent developments in computerized measurement are summarized by placing them in a four-generation framework in which each generation represents a genus of increasing sophistication and power as follows: (1) computerized testing--administering conventional tests by computer; (2) computerized adaptive testing--tailoring the difficulty or contents of the next piece presented or an aspect of the timing of the next item on the basis of examinees' responses; (3) continuous measurement--using calibrated measures embedded in a curriculum to continuously and unobtrusively estimate dynamic changes in the student's achievement trajectory and profile as a learner; and (4) intelligent measurement--producing intelligent scoring, interpretation of individual profiles, and advice to learners and teachers by means of knowledge bases and inferencing procedures. The suggested framework may contribute to discourse in the field and facilitate communication about the rapidly developing issues. (Contains 5 tables, 2 figures, and 158 references.) (Author/SLD)

****
*   Reproductions supplied by EDRS are the best that can be made   *
*   from the original document.   *
****

RR-88-35

ED 395 000

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it
☐ Minor changes have been made to improve reproduction quality
- Points of view or opinions stated in this document do not necessarily represent official OERI position or policy

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
H. I. BRAUN

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

RESEARCH REPORT

THE FOUR GENERATIONS OF COMPUTERIZED EDUCATIONAL MEASUREMENT

C. Victor Bunderson
Dillon K. Inouye
James B. Olsen

Educational Testing Service
Princeton, New Jersey
June 1988

TM024996

BEST COPY AVAILABLE

THE FOUR GENERATIONS OF COMPUTERIZED EDUCATIONAL MEASUREMENT

C. Victor Bunderson
Educational Testing Service

Dillon K. Inouye
Brigham Young University

James B. Olsen
WICAT Systems, Inc.

This chapter appears in R. L. Linn (Ed.), 1988
Educational Measurement (3rd ed.). New York: Macmillan

Copyright © 1988. Educational Testing Service. All rights reserved.

The Four Generations of Computerized Educational Measurement

Abstract

Educational measurement is undergoing a revolution, due to the rapid dissemination of information-processing technology. One of the most notable aspects of that revolution is the rapidity with which it has come upon us. It is perhaps inevitable that the recent growth in power and sophistication of computing resources and the widespread dissemination of computers in daily life have brought about irreversible changes in educational measurement.

Recent developments in computerized measurement are summarized by placing them in a four-generation framework, in which each generation represents a genus of increasing sophistication and power.

Generation 1. Computerized testing (CT): administering conventional tests by computer

Generation 2. Computerized adaptive testing (CAT): tailoring the difficulty or contents of the next piece presented or an aspect of the timing of the next item on the basis of examinees' responses

Generation 3. Continuous measurement (CM): using calibrated measures embedded in a curriculum to continuously and unobtrusively estimate dynamic changes in the student's achievement trajectory and profile as a learner

Generation 4. Intelligent measurement (IM): producing intelligent scoring, interpretation of individual profiles, and advice to learners and teachers, by means of knowledge bases and inferencing procedures

While acknowledging the obvious pitfalls associated with proposing a framework in a developing field, the authors hope that the suggested framework will provide an *ad interim* contribution to the field's universe of discourse and facilitate communication about the rapidly developing issues.

THE FOUR GENERATIONS OF COMPUTERIZED EDUCATIONAL MEASUREMENT

Contents

Introduction 1

- Purpose of this chapter 3
- Defining dimensions of the four generations 4

The first generation: Computerized testing (CT) 26

- Definition 26
- Advances in test administration in the CT generation 27
- Creating tests and items by computer 40
- Research issues for the CT generation 41
- Engineering design issues 45
- Experimental design issues 47
- Scientific issues 50

The second generation: Computerized adaptive testing (CAT) 52

- Definition 52
- Examples of computerized adaptive tests 53
- Adapting on the basis of item parameters 56
- Current computerized adaptive testing systems 60
- Advantages of computerized adaptive tests 68
- Research problems with computerized adaptive testing 72

The third generation: Continuous measurement (CM) 77

- Definition 77
- Examples of partial CM systems 81
- Mastery assessment systems as continuous measurement 95
- The role of learner profiles in the continuous measurement generation 106
- Research issues in continuous measurement 114

The fourth generation: Intelligent measurement (IM) 116

- Definition 116
- Three potential contributions of IM to test administration 119
- Intelligent tutors: A converging or discontinuous line of development? 125
- Complications of artificial intelligence: Future generations 129

Summary 131

Concluding thoughts 134

- Generational enhancements in powers of observation 134

References 136

THE FOUR GENERATIONS OF COMPUTERIZED EDUCATIONAL MEASUREMENT

C. Victor Bunderson
Educational Testing Service

Dillon K. Inouye
Brigham Young University

James B. Olsen
WICAT Systems, Inc.

INTRODUCTION

Educational measurement, the specification of position on educationally relevant scales, is undergoing a revolution, due to the rapid dissemination of information-processing technology. Because the process of measurement is labor intensive, it is not surprising that the exponential increase in our capacity to do work should revolutionize educational measurement, making it possible for both the psychometrician and the consumer of psychometric services to do routinely what was previously impossible.

One of the most notable aspects of the revolution is the rapidity with which it has come upon us. Although other major innovations in education, like writing and printing, took centuries and even millennia

---

Note: The authors acknowledge with gratitude the assistance of Robert Linn, the general editor, and Bill Ward, Howard Wainer, George Powell, Garlie Forehand, and Randy Bennett, of Educational Testing Service, who reviewed earlier versions of the manuscript and made suggestions of substance that led to significant improvements. Myrtle Rice and Jeanne Inouye provided excellent editorial assistance. Kevin Ho coordinated production details with the two authors in Utah. Bobbi Kearns, Alice Norby, and Joyce Thullen were excellent under pressure in manuscript production and revisions.

to become the common possession of everyone, the distribution of computing resources has occurred within decades. A measure of the rapidity with which computers have been adopted by educational measurement is seen in the fact that the previous edition of Educational Measurement, published in 1971, did not include a chapter on the subject. This was true despite the fact that a number of promising early experiments had been conducted, that computers were widely used in test scoring, and that a book had been published that included the words "computer assisted testing" in the title (Holzman, 1970).

The computer revolution has been marked by the growth in power and sophistication of computing resources. The computing power of yesterday's mainframes is routinely surpassed by today's supermicros. Yesterday's ENIAC computer, which filled an entire room, was less powerful than the current generation of microcomputers, which fit on a desktop. Computers that are not sophisticated or powerful enough for educational measurement can now be easily connected to computers that are.

The computer revolution has also been marked by the widespread dissemination of computers in daily life. Yesterday, computing power was the exclusive possession of a few; today it is available to everyone. Yesterday, only the cognoscenti knew about computers and their related arcana; today one is embarrassed not to be computer literate. The recent Commission on Excellence in Education (Gardner, 1983) formally acknowledged the ubiquity and importance of computers in our society by branding American students "illiterate" in their knowledge of computers. This unprecedented characterization conveys the expectation that everyone should be familiar with computers. It signals one of the largest general education (and reeducation) tasks in history.

Perhaps inevitably, these changes in the power and distribution of computing resources have wrought irreversible changes in educational measurement. No evidence of the revolutionary character of these changes is stronger than the announcement, in recent years, of large-scale computerized measurement projects. The armed forces are developing a computerized version of the Armed Services Vocational Aptitude Battery (Green, Bock, Humphreys, Linn, & Reckase, 1982). Educational Testing Service has announced a major commitment to new priorities which will include the use of computerized measurement systems to better serve individuals (Ward, 1986), and it has implemented operational systems. The State of California is developing a computerized prototype of its future Comprehensive Assessment System (Olsen, Inouye, Hansen, Slawson, & Maynes, 1984).

The significance of these large bellwether projects is that they show the direction in which the field is moving. They are milestones marking the transition from the classical era in the field of educational measurement to the beginning of another era, a transition that will affect every psychometrician and consumer of psychometric services.

PURPOSE OF THIS CHAPTER

In what follows, we shall attempt to summarize recent developments in computerized measurement by placing them in a four-generation framework (Inouye & Bunderson, 1986; Inouye & Sorenson, 1985). In this framework, each generation represents a genus of increasing sophistication and power. We suggest that the framework be used as temporary scaffolding, to be discarded when more useful and powerful representations are built. Despite the obvious pitfalls associated with proposing a framework in a developing field, we hope that our suggestion of a four-generation framework will provide an *ad interim* contribution to the field's universe of discourse, one that will facilitate communication about the rapidly developing issues. Our nominees for the four generations are

> *Generation 1. Computerized testing (CT)*: administering conventional tests by computer
>
> *Generation 2. Computerized adaptive testing (CAT)*: tailoring the difficulty or contents of the next piece presented or an aspect of the timing of the next item on the basis of examinees' responses
>
> *Generation 3. Continuous measurement (CM)*: using calibrated measures embedded in a curriculum to continuously and unobtrusively estimate dynamic changes in the student's achievement trajectory and profile as a learner
>
> *Generation 4. Intelligent measurement (IM)*: producing intelligent scoring, interpretation of individual profiles, and advice to learners and teachers, by means of knowledge bases and inferencing procedures

We will now present defining attributes of computerized educational measurement and discuss the four generations, suggesting some advantages, challenges, and immediate opportunities for research.

DEFINING DIMENSIONS OF THE FOUR GENERATIONS

Computerized educational measurement is a subfield of educational measurement that is formed by the intersection of educational measurement and the technology of computer delivery systems.

Computerized educational measurement is therefore that area formed by bringing educational measurement and computing resources into relationship with each other.

Educational Measurement

Definition. Educational measurement is the process of specifying the position or positions, for educational purposes, of persons, situations, or events on educationally relevant scales under stipulated conditions. This definition is given to provide a framework for six categories along which types of educational measurement can differ and which enable us to contrast the attributes and properties of the four generations.

Educational measurement as process. Educational measurement is a process composed of several subprocesses, some occurring in parallel, some in series. Major processes are (a) test development, including development of test specifications and candidate items, pretesting, and combining of selected and revised items into tests; (b) test administration, including obtaining, scoring, reporting and interpreting responses; and (c) test analysis and research, including equating, linking, validating, and analyzing for differential item functioning and group differences.

Computers are now having a major impact on all three classes of educational measurement processes because of their ability to do work at electronic speeds. However, this chapter will narrow its focus to the subject of test administration; it will discuss obtaining examinee responses, scoring them, recording them for later use, reporting and interpreting the results, and giving prescriptive advice.

Specification of position. Educational measurement specifies a position, or positions, along educationally relevant scales. Specification can be static, measuring the position of a person, situation, or event at one particular time, or dynamic, measuring changes in position over time. Precision, reliability, power, and efficiency are other dimensions along which specification can differ.

The essential difference between static and dynamic measurement can be seen by analogy to the physical sciences. Early physicists were limited to an understanding of statics, the properties of objects in their motionless states. In contrast, later physicists invented tools that helped them understand dynamics, the properties of objects in motion. When early physicists studied a weight suspended from a spring, they could only measure the distension of the spring when it stopped moving. In contrast, later physicists could understand both spring and weight as parts of a dynamic system in oscillatory motion.

The distinction between static and dynamic measurement in the physical sciences is analogous to the distinction between outcome and process in the social sciences. The measurement of achievement in U.S. public schools is an example of static measurement, because the purpose of measurement is to specify the state of learners with respect to achievement variables at single moments in time, usually at the beginning or end of the year. On the other hand, the measurement of growth in individual achievement as a function of instruction is an example of dynamic measurement, because the purpose of measurement is to describe changes in the learner over time.

If static measurement is the specification of a point, or points, in an educationally relevant measurement space, dynamic measurement is the specification of a trajectory, or path of points, over time. If a point defines a position along a relevant scale, then a trajectory defines changes in position over time.

A given trajectory of a person can be essentially linear, specifying uniform translation of position over time, or it can be curvilinear, specifying not only change in position but also change in the rate of change. Linear translation is analogous to constant velocity; curvilinear motion over time is analogous to acceleration. The first and second generations of computerized measurement usually deal with static measurement, and the third and fourth generations usually deal with dynamic measurement.

Educational purposes. Traditionally, the principal purpose of educational measurement has been to assist educational decision making by providing information about the position of a group or an individual along educationally relevant scales. Measurement has typically served two constituencies, institutions and individuals.

Historically, educational measurement has been used primarily for institutional purposes. Institutions use measurement to improve their admissions and placement decisions, to assess the achievement of educational goals, to evaluate personnel and programs, to evaluate organizational entities, and to motivate students. The traditional uses of educational measurement to serve individuals have included guidance and counseling of people, based on achievement, ability, aptitude, or interest test scores; monitoring of individual progress; and assistance in instructional decision making.

In addition to those uses of measurement for individuals that imply more formal and standardized application, there are a myriad of informal uses for unstandardized educational measurement by teachers, learners, and administrators. It has not yet proved cost effective, however, to provide educational measurement conforming to the high standards of development, administration, analysis, and use established by professional organizations and applied in admissions testing, research, and major summative evaluation projects.

The field is therefore open for serving individual purposes with new kinds of excellent educational measurement. In addition to guidance and counseling of individuals, educational measurement can be used to monitor learner trajectories in well-defined educational measurement spaces. It can be used to diagnose problems in velocity and acceleration and provide information for timely instructional intervention. Good educational measurement can provide data for profiling the characteristics of individuals and their progress in an achievement space. It can also guide the interpretation of these profiles and lead to prescriptive advice for individuals based upon their learner profiles and achievement trajectories.

Monitoring, profiling, and interpreting are individual purposes of educational measurement that are closely linked to instruction. Other instructional activities that could be more closely linked to measurement include selecting appropriate scope and sequence, providing nontrivial instructional guidance within that scope and sequence (as would an excellent coach), and recommending practice on exercises at the appropriate level for each individual.

The shift in emphasis from institutional purposes to individual purposes characterizes the distinction between the first two and the last two generations of educational measurement. This distinction is closely related to the shift from static to dynamic measurement.

Persons, situations, and events. The objects of educational measurement are persons, situations, and events. It is commonplace in the social sciences to see behavioral events (B) arising as a function of the interaction between person (P) and situation (S) variables; in other words, $B = f\ (P,S)$. Educational events, like learning, may also be seen under a similar functional rubric: learning (B) = f (aptitude (P), treatment (S)).

Although global situations can be measured (e.g. "educational climate"), the important subset of situations addressed in this chapter is specially designed and calibrated tasks used to specify position along educationally relevant scales. In these cases, according to the behavior formula cited, a task situation, S (the test item), is presented and a behavioral event, B (the examinee's response), is observed and scored, in order to infer or measure an examinee's relative position on a scale (P).

The standardized task situations in the first two generations are test items. Usually the observed behaviors are given a binary score, right or wrong, but responses to more complex tasks might be scored holistically with a graded numerical score. A series of items, or situational tasks, which can be shown empirically to vary along a single dimension, are calibrated, and their position is specified along an inferred scale of a latent (unobserved) trait. In the third generation, the situational tasks become more complex, consisting of multiple responses and more realistic, worldlike simulations. These situations can also be measured and calibrated.

Aggregated data on persons or tasks are often used to make inferences about educational programs or about constructs thought to explain group differences on the scales. The scales provide a model that focuses on some dimension of the knowledge or skill domain and the positions of groups on these dimensions.

A broad class of events can have educational significance. Some of these, like participation or nonparticipation in a particular activity, might merely be noted and become a part of an educational record. Other events are significant because of the time they require. The measurement of time intervals is enormously enhanced by computerized measurement.

For each class of persons, situations, or events, the advent of computerized measurement allows measurements that were previously impossible. This is true in at least two ways. In the first, or practical, sense, some objects previously unmeasured because of lack of money, time, and expertise can now be measured. For instance, expensive individually administered intelligence and aptitude tests can now be administered more inexpensively to more individuals by computer under the supervision of paraprofessionals. Other examples include case studies and simulations, previously administered manually or not at all, and frequent measures to produce learning trajectories.

A second way in which the previously impossible becomes possible is due to changes in the operating capabilities of the measurement delivery system. An example of this is seen in mental chronometry. Here, the chronometric properties of certain mental events, like the relative speeds of mental rotation of geometric figures, can be measured and recorded, scored, and interpreted. Other examples are the automatic processing of types of responses, digitized vocal responses, and movements of a joystick or a mouse.

Educationally relevant scales. Educational measurement is defined here as the specification of position along educationally relevant scales or dimensions. The dimensions of measurement spaces are always constructs, conceptual inventions, that are imposed on the persons or tasks being measured. They do not inhere in the objects themselves. Even in the physical sciences, constructs like weight, mass, and energy had to be invented before measurement could occur.

The constructs of education and the social sciences differ from those of the physicochemical sciences in both their degree of theoretical interrelationship and their empirical grounding. For example, although in physics $1^{\circ}C$ is theoretically related to mass and velocity through the formula $1/2\ \underline{mv}^2$, in the social sciences, no such network of interrelationships has been uncovered between, say, IQ scale points and mathematics achievement. In education, the lack of ratio scales, the lack of agreement on constructs, and the complex dimensionality of the area measured are factors that have posed severe difficulties in finding theoretical linkages.

Stipulated conditions. The final item in the definition of educational measurement refers to the conditions under which measurements are taken. As Cassirer (1923) has argued, the value and usefulness of any measurement are dependent upon specification of the conditions under which the measurement is made. When conditions differ, the meaning of two or more measurements can differ. The degree to which measurement conditions can be specified is the degree of control. To the extent that measurement conditions can be controlled, we may say that they are standardized. The threat of extraneous sources of variation is then minimized, and the conditions are made replicable. The four generations of computerized measurement add important new contributions to the control and standardization of measurement conditions, making possible comparisons between measurements of objects and events previously thought to be incommensurable.

The Computerized Delivery System

Delivery system, work, and information technologies. The second set of dimensions that define computerized educational measurement refers to variations in the work capacities of delivery systems. The process of specifying position in an educationally relevant space requires a combination of theory, methods, and work. Theory is necessary to invent the constructs that define a measurement space. Methods are necessary to improve the quality, that is, the reliability and validity, of measurement. Work is necessary to process the information needed for the specification of position. We now turn to computing resources that supply the work necessary for information processing.

BEST COPY AVAILABLE

Administering, scoring, recording, reporting, and interpreting are labor-intensive processes. The rapid deployment of computing resources insures the widespread capability to do this work at electronic speeds for increasingly lower costs.

In this chapter, we shall refer to the computing resources provided by modern information technologies as the delivery system or, more simply, as the computer. The delivery system includes the hardware, software, testware, and human expertise necessary to deliver the intended instruction or measurement. Technology is not limited to hardware; it refers, more generally, to the application of knowledge.

Hardware. The hardware components of a single workstation of a modern computerized measurement system typically include:

1. A single computer, possibly joined to others through a local area network or a long-distance communication line. The workstation could also be a terminal for a multiuser computer.
2. Sufficient memory for the applications intended
3. Mass storage capacity, such as floppy disks, fixed disks, or videodiscs
4. A response input device or devices
5. Display devices for text and graphics and, sometimes, audio
6. A printer
7. Data communications to a central site

There are a large number of permutations and combinations of these essential hardware elements, as well as many enhancements to this basic system.

Software. To the hardware must be added the following essential software components:

1. An operating system with device drivers and utilities to harness and coordinate the resources of the delivery system
2. Applications software (testware), for administering, scoring, recording, reporting, and, in some cases, interpreting the results.

Software is the intelligence that channels and directs the work of the delivery system. Computerized-testing software has advanced considerably, and new software has been implemented in a variety of delivery systems for computerized and computer-adaptive tests (CAT). Several new item calibration programs are also available for CAT.

Among the more significant developments in software are those associated with knowledge-based (artificially intelligent) computing. Methods of developing knowledge bases and procedures for querying these knowledge bases open the prospect for the fourth generation of computerized testing, intelligent measurement. The software needed for this generation combines advances in both computerized testing algorithms and knowledge-based computing.

Implementation policies and strategies. The bitter lesson of many attempts to promote technological revolution is that revolutions only partially depend upon advances in hardware and software. The rate of revolution depends on people. Unless those responsible for sponsoring and maintaining the delivery system learn how to become effective change agents, successful transition from a print-based culture to an electronic one is unlikely. Technology is the application of knowledge, and knowledge has an important personal component. Because people carry their knowledge in their bodies, the transfer of knowledge occurs one person at a time.

The implementation of a successful computerized testing operation, therefore, requires a thorough, tested set of policies and procedures for training over a sufficient period of time. These procedures can insure that the computerized testing system is implemented in such a way that it achieves benefits and maintains conditions for validity and equity in its use.

Current state of the art. Our discussion of the essential hardware, software, and implementation policies as essential components of the delivery system has prepared us for a discussion of the state of the art in each of the component areas. Because the components of modern delivery systems are changing so rapidly, any attempt to catalog the current state of the art will quickly be outdated. Today's state-of-the-art devices might become exhibits in tomorrow's museum of antiquities. We discuss five generic kinds of work done by the delivery system that will persist, even when hardware and software become obsolete. For each kind, we illustrate trends of development.

Five kinds of work. Discussed next are five dimensions of work along which delivery systems, or their components, can differ. Each dimension represents a different kind of work that has its analog in human performance. Along each of these dimensions, technology has exponentially increased the amount and kinds of work that an individual can do. The five dimensions are: sensing, remembering, deciding, acting and communicating. It is in the particular combination of these five kinds of work that delivery systems in the four generations differ.

Sensing. Input devices do the work of sensing. They pick up information from the examinee or the environment, encode it as symbols, and transmit it to the system for interpretation and response. Input devices are evolving rapidly from keyboards to window-type interfaces, pull down menus, icons, and mouse. Also available are touch screens, joysticks, and trackballs. Input by means of keyboards gives an advantage to examinees who have had previous experience, such as touch-typists. The expanded use of voice recognition and other methods of input might equalize advantages.

Remembering. Memory devices do the work of remembering stored information. They allow the system to remember the step-by-step sequence of operations it is to perform and the instructions and data it is to use. As with humans, memory makes it possible for the machine to recognize signals, decode stored instructions, record data, adapt to records of past experience, and organize data into structures so that it can process these higher order units.

The memory capacity of most modern delivery systems is evolving rapidly. Most microcomputer workstations now have from 1/2 to 2 megabytes of random access memory. Future workstations will use even larger amounts of random access memory. The early, expensive, mass-storage devices are being replaced by inexpensive, high-density, magnetic and opto-electronic devices. Hard-disk storage exceeding 100 megabytes per workstation is becoming more common. Compact disk read-only memories, Write Once Read Many optical disks, and videodiscs will soon allow gigabytes of storage per workstation.

Deciding. Microprocessors that do the work of deciding perform the calculations necessary to make the decisions. This includes processing inputs, computing, and making decisions based upon information in memory. In the system, the work of deciding is done by the central processing unit(s). It performs the mathematical and logical operations required to make a decision. It also controls the operation of the machine by activating the computer's other functional units at appropriate times.

Most delivery systems of the future will utilize microprocessors that handle at least 16 to 32 bits at a time at speeds of from 6 to 50 million cycles per second. The evolution is toward larger information-handling capacities at higher speeds. The current state of the art is represented by 32-bit microprocessors, which have a full 32-bit architecture, a full 32-bit implementation, and a 32-bit data path (bus) to memory. Some technology writers predict that microprocessor performance will eventually exceed that of all but a few of our current mainframes.

Acting. Output devices execute the decisions made by the system. This work includes activating output devices that send information, turn on motors, switch lights, display the next test item, and position and activate mechanical devices. The work of acting allows the computer to change the environment or to control devices external to the system. It also allows the computer to communicate with people and with other machines. The most important subcategory of acting for the purposes of computerized measurement involves controlling display devices for the text, images, and audio used in testing situations.

Output devices are also undergoing rapid evolution in performance, price, and variety. Visual displays have improved tremendously since the days of the teletype. Soft-copy displays have become more and more prevalent, most of them in the form of cathode-ray tube (CRT) displays. The CRT is still the display of choice for most testing applications, although its competitors, liquid crystal, electroluminescent, and plasma displays, have made impressive gains and are gaining greater currency.

Communicating. In addition to the types of work listed, which, in combination, make a given delivery system more or less powerful, linking computers can also increase the amount and sophistication of work performed. Here, too, the cost of work devices relative to performance is decreasing. Local area networks make possible the linking of multiple workstations for individualized testing applications. Long-haul networks make possible the distribution of upgraded norms and experimental items from central test-development sites and the collection of statistics from distributed sites. Future generations of computerized testing will require linked workstations to make use of the additional capabilities afforded by these developments, such as group-interactive tests for assessing team performance.

Computerized measurement systems can replicate many kinds of work hitherto done by humans. Some examples from which large savings of time and energy have resulted include the scoring of tests, the searching of large files to retrieve records or test items, the computing of statistics, the processing of text, and the keeping of records. Devices that do such work can be widely disseminated to increase exponentially the work available for educational measurement.

Table 1 summarizes our discussion to this point. It shows differences among the four generations of computerized measurement, based on differences in computer sophistication and in the six defining attributes of educational measurement. The generations have many superficially similar elements, but, just as a Model T differs from a modern automobile, the generations differ from each other. These differences affect the efficiency, speed, convenience, accuracy, and power of educational measurement.

Table 1

Features of Four Generations of Computerized Educational Measurement

Generation	Computerized Testing (CT)	Computer-Adaptive Testing (CAT)
Computerized delivery system features	Computer-controlled administration; rapid scoring and reporting; new display and response types; mass storage for displays and item banks, network communications	Same as CT Fast floating-point calculations for adaptive algorithms
Scientific foundations	Varied, but usually classical test theory	Item response theory & related advances
Educational measurement functions		
Processes	Administering, Scoring, Recording, Reporting	
Specification of position	Static (usually)	Static (usually)
Educationally relevant purposes	Institutional (usually)	Institutional (usually)
Scales	Varied (can be informal)	Unidimensional for IRT-based tests; evolving
Educational objects	Persons Situations Events	Persons Standardized tasks Events
Degree of control	High for display and responses	More adaptive control than CT

Table 1 (continued)

Features of Four Generations of Computerized Educational Measurement

Generation	**Continuous Measurement (CM)**	**Intelligent Measurement (IM)**
Computerized delivery system features	All of CAT features Computer-aided education features	All of CM features Knowledge-based inferencing
Scientific foundations	Extensions of IRT Valid construct specifications Learner profiles Implementation design	Models of expert knowledge-- scoring expertise, profile interpretation, teaching expertise
Educational measurement functions		
Processes	Same as CAT plus more interpretation	Same as CM plus sophisticated interpretation
Specification of position	Dynamic (static possible)	Same as CM
Educationally relevant purposes	Individual (institutional possible)	Same as CM
Scales	Multidimensional Composite	Same as CM when needed
Educational objects	Persons Reference tasks Events	Same as CM
Degree of control	CAT plus control over instruction	Same as CM, but much control can be given to user

Summary of Computerized Delivery System Features

The computer capabilities of the four generations have much in common. All four permit computer-controlled administration, rapid scoring and reporting, new display and response types, mass storage for displays and item banks, and network communications. The first generation does not require a fast floating-point processor for the item-by-item calculations required by some adaptive algorithms in the second generation. The third generation has, in addition, the computer-controlled features of display, response entry, and processing needed in computerized instruction. In the continuous measurement generation, testing disappears into the fabric of instruction and measurement becomes unobtrusive. Artificial intelligence allows drawing inferences from knowledge bases to provide more sophisticated scoring, interpretation, and advice in the fourth generation.

The scientific foundations of measurement differ among the generations. The first generation is frequently characterized by the use of classical test theory or by the lack of underlying psychometric theory. Individuals familiar with interactive computing frequently implement tests in an *ad hoc* manner. They are either unaware of, or unconcerned with, measurement issues such as validity, reliability, and equating from paper-and-pencil or individually administered versions to the computer version. The face validity of a new simulation-like test is often seen as sufficient.

The scientific bases of the second and higher generations are more advanced and are the subject of much current research. The second generation has prominently featured adaptive algorithms based on item response theory (IRT), and it is, consequently, limited to situations in which the assumption of unidimensionality of the underlying scale can be demonstrated, although current work could change this. The third generation will not reach its full potential until there are extensions of IRT or new psychometric theories to allow entities other than items to be calibrated.

New developments in psychometric theory are a necessary, but not sufficient, scientific basis for the continuous measurement generation. Valid construct specifications of underlying scales and cognitive components are necessary for the successful use of calibrated item clusters and tasks. Also necessary, but coming later during the evolution of CM, is the measurement of learner profiles representing different learning abilities, styles, and preferences. All of these advances will fail unless implementation design principles and techniques are developed. Users will need careful and extensive in-service training, because new roles and traditions will have to evolve to take advantage of continuous measurement. A research basis for implementation design is thus a critical task for applied science. The fourth generation will introduce new scientific foundations including models of expert knowledge to accompany computer applications. Promising applications will include the knowledge to score complex tasks, the professional knowledge to interpret profiles, and the knowledge of teaching experts capable of using data from continuous measurement and from learner profiles to provide prescriptive advice to learners and teachers.

Summary of Educational Measurement Functions

The generations do not differ extensively on test-administration processes. The main distinction among them is in the extent to which the computer system is programmed to provide interpretation to the user, a major function usually reserved for a counselor or a teacher acting as counselor. Some interpretation by computer is possible in the continuous measurement generation, but knowledge-based computing with the programmed expertise of the teaching expert is necessary before sophisticated interpretation, analogous to the human teacher or counselor, will be possible.

It is useful to point out that the first two generations usually deal with static measurement, whereas the last two emphasize dynamic measurement. This fact is closely related to their educational purposes. The first two generations primarily serve institutional functions, because the psychometrically sophisticated tests implemented on computers so far are usually variations of current tests used for institutional purposes. The third and fourth generations emphasize individual educational purposes.

The measurement scales among the generations vary in psychometric sophistication. In the first generation we might have varied measurement scales. They could be informal and ordinal, nominal, or interval scales. The second generation requires a unidimensional equal-interval scale for tests based on IRT. The third generation requires that we also deal with the multi-dimensionality inherent in learning any complex knowledge domain. It is necessary to develop composite scales to provide learners with reports of overall progress on more than one underlying scale. For providing advances in scoring and interpretation in the process of learning, intelligent measurement requires all of the scale sophistication of the CM generation. Intelligent measurement can also be used to enhance the scoring of first- and second-generation tests, in which case it might revert to simpler scales.

Item response theory has provided the major scientific advancement for developing educationally relevant scales. It enables us to obtain scale values for both persons and tasks on the underlying single dimension. Item response theory scaling is widely used in second-generation applications and is possible on the first generation with a nonadaptive IRT test. The calibration of tasks more complex than multiple-choice items is one of the attributes which defines the third generation. By extending calibration methods from items to the more complex tasks used in instruction, it becomes feasible to track individuals in an educationally interesting growth space.

The degree of control refers to the stipulated conditions under which measurement can be standardized. The first two generations permit great control over the display and sequencing of visual and auditory stimulus materials, the form of response, and the timing of responses. The third generation introduces additional control over instructional events and deemphasizes the distinction between instructional and testing events. The fourth generation could be used to introduce another degree of control: control of the process of instruction by the learner. Intelligent measurement thus poses some problems in standardization, due to increased user control over instructional options.

This concludes our presentation of the defining context in which computerized educational measurement may be viewed. In the next sections, we consider in some detail the unique promise and problems inherent in each generation.

THE FIRST GENERATION: COMPUTERIZED TESTING (CT)

In the first generation of the computerization of any human activity, users tend to automate familiar but manually time-consuming processes. Later, having become more familiar with the capabilities of the computer, they begin to see ways in which computer power can be used to perform previously impossible or even unimagined tasks. This pattern can be seen in the evolution of computerized measurement. In the first generation, computerized testing (CT), computers are used to automate familiar measurement activities. The CT generation began with the translation, or conversion, of familiar paper-and-pencil tests, usually group administered, to a computer-administered format. The CT generation also includes the development of new nonadaptive tests, similar to manually administered tests, but more efficient in utilizing computer capabilities for administration. In nonadaptive tests the number of items, their sequence, content, and timing, do not depend on examinees' responses in any way.

DEFINITION

The first, or CT, generation is defined as the translation of existing tests into a computerized format, or the development of new, nonadaptive tests that are similar to manually administered tests, but utilize computer capabilities for all or most test-administration processes. The CT generation tests usually report a static position on an ordinal scale, and the scores are used for institutional purposes far more frequently than for individual purposes.

First-generation tests now constitute the largest set of exemplars of computerized measurement, and they will continue to proliferate. They do not require complex algorithms and psychometric models or the more sophisticated computer requirements of the second and higher generations. They are generally used for familiar purposes that do not require dramatic role shifts on the part of users (such as in the new teaching and learning roles found in the third generation).

ADVANCES IN TEST ADMINISTRATION IN THE CT GENERATION

Although it introduces some new problems, computerized testing advances many of the processes of test administration. In this section, we will review the improvements and problems associated with these test-administration procedures: presenting item displays, obtaining and coding responses, scoring, reporting and interpreting results, and collecting records at a central site.

Presenting Item Displays

Greater standardization. Administration of computerized testing introduces precise control over item displays. The timing of the displays can be precise, as can control over what the examinee sees or hears.

Computer-administered testing permits test administration conditions, directions, and procedures to be completely standardized in ways not possible with manually administered tests. Computerized test directions are always the same, no matter how many times the test is administered and no matter how many different locations or test administrators are involved. Some instructions are precluded, and others can be enforced. Instructions like "Do not turn the page until I give the signal" are not needed. The computer can rule out peeking by controlling the displays.

Greater standardization might, however, imply greater difficulty in adjusting testing conditions to meet local needs. The computer can be programmed to be very resistant to alterations in test administration conditions, or it can be programmed to give the examiners flexibility to alter testing conditions in certain prescribed ways, such as breaking a test into two different time intervals or restarting at the appropriate point with a review of the instructions for the last subtest.

Improved test security. Computerized testing also provides for increased test security. There are no paper copies of the tests or answer keys to be stolen, copied, or otherwise misused. Computer-administered tests can also include multiple levels of password and security protection to prevent unauthorized access to the testing materials, item banks, or answer keys. Test displays and item keys can be encrypted to prevent unauthorized printing or copying and the test items and associated answer keys can be randomly resequenced, if necessary, so that a student cannot follow another examinee's screen.

Enriched display capability. A printed test has obvious display strengths and limitations. It can present text and line drawings with ease. At greater cost, photographic illustrations can be presented. A printed test cannot provide timing, variable sequencing of visual displays, animation, or motion. Audio devices can be used for standardized audio presentations associated with a printed test, but a trained administrator must deliver the audio in a group-paced mode or to the examinees one by one.

In the CT generation, a visual display device replaces the printed page display. The quality of the display varies with the resolution, graphics circuitry, bandwidth, and memory available for storing display data. As an example of what can be accomplished with good display features, Druesne, Kershaw, and Toru (1986) have been able to implement the CLEP artistic judgment test, which uses photographic illustrations, on an IBM personal computer equipped with an advanced graphics board, a high-resolution color digital monitor, and sufficient storage for the digitized photographic images. Even more advanced displays are possible, which equal the resolution of a printed photograph. Video images can be stored on a videodisc, which permits random access to single video images or short motion sequences and can provide for the dynamic overlay of text or graphics on the video. For example, ETS has developed an interactive video test (CT generation) using a computer-controlled videodisc with graphics overlay in the areas of medical certification (podiatry), and English as a Second Language (Bridgeman, Bennett, & Swinton, 1986).

In general, computerized displays sacrifice some image resolution for greater flexibility and control over the presentation of text, graphics, animation, motion, audio, and video. System costs go up with the addition of color, audio, graphics resolution, and videodisc. The user is thus faced with a cost-capability trade-off when choosing a lower or higher resolution display screen and the presence or absence of audio and video. This trade-off may cause problems with CT, because the resolution of the display screen determines how much of an existing test item can be shown at one time without scrolling or paging and how accurately line graphics or photographs can be reproduced.

New item types. New item types can be developed using advanced display capabilities. Such items are part of the first generation, so long as they are not presented adaptively, are not part of a continuous measurement system embedded into a curriculum, or do not utilize intelligent advice or scoring. WICAT System's Learner Profile, a battery of 45 CT and CAT tests covering a variety of learning-oriented aptitude and preference dimensions, provides numerous examples. Among its item types are gestalt completion items, in which more and more image detail is unfolded until an examinee recognizes a picture; animated displays that test visual concepts and memory for spatial sequences; accelerating object displays to test perceptual speed; and individually-administered audio presentations using earphones and computer-controlled digitized audio to test auditory digit span.

Equivalent scores with reduced testing time. For the majority of items provided in current standardized achievement and psychological tests, computer-administered versions offer the promise of significant reductions in test-administration time. In two different research studies using subtests and items from the California Assessment Program, Olsen, Maynes, Slawson, and Ho (1986) found that computer-administered and paper-and-pencil tests produced the same statistical test-score

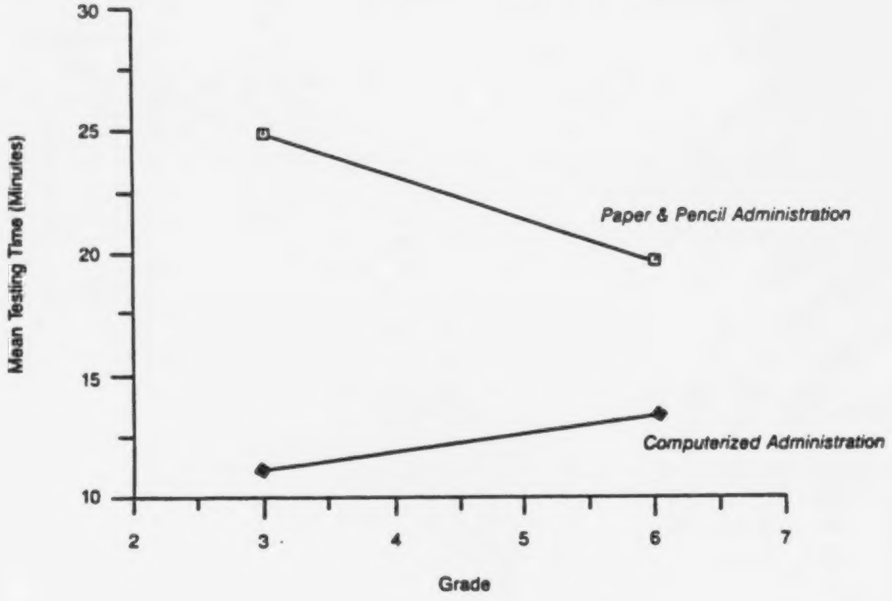
BEST COPY AVAILABLE

distributions (means, standard deviations, reliabilities, and standard errors of measurement), but the computer-administered test required less administration time. Like the majority of standardized achievement tests, the California Assessment Program uses a computer-scannable answer sheet and test booklet. Gridding answers on a separate sheet appears to take more time than responding to a screen display by using a keyboard. Figure 1 shows that the computer-administered tests were completed by examinees in significantly less time than the paper- and-pencil tests. Other studies have shown similar patterns, but the issue is not simple (Calvert & Waterfall, 1982; Watts, Badeley & Williams 1982). The effect for younger children is shown in Figure 1 to be larger, implying that the use of answer sheets is less natural and, consequently more likely to be a disadvantage for children who lack test-taking experience.

Obtaining and Coding Responses

The increased speed of some computer-administered tests over similar manually-administered tests is at least as much a function of improvements in obtaining and coding responses as of increased control over the presentation and pacing of displays. A mark-sense sheet commonly used in manual testing requires the examinee to code each response by associating it with the item number and then marking one of several bubbles. Visually matching item numbers and alternative numbers or letters takes time and produces errors. By presenting only one item per screen, the computer automatically matches responses with the item number, makes alternatives visually immediate, and reveals the

FIGURE 1


30
25
20
15
10
Mean Testing Time (Minutes)
2
3
4
5
6
7
Grade
Paper & Pencil Administration
Computerized Administration

selections for immediate verification. Convenient features for changing answers can replace time-consuming erasing on printed answer sheets. Computers can also administer item formats other than limited-choice items. Constructed responses involving numbers or formulas entered on a keyboard are quite easy to interpret unambiguously. Short answers involving keywords require more sophistication, but, even without natural-language-processing techniques, highly accurate coding can be obtained. One method of achieving good results is to obtain and sort a large collection of pretest answers. With this data, keyword processing with misspelling tolerance can be used with a range of acceptable answers. These methods can, with high accuracy, produce correct, partially correct, or incorrect item coding.

Computers also facilitate other response formats, such as pointing out and marking words in text or parts of pictures or drawings. Such forms of response entry have been possible in computer-aided instruction and in personal computing for many years. These response forms, increasingly easy and familiar to a population of examinees whose computer literacy grows yearly, can be standardized and automated to become an important part of the CT arsenal of item types. To compare computer-based response formats with their paper-and-pencil equivalents, we need to design studies in which paper-and-pencil and CT versions of the same test are administered. Thus, coding of short answers, flexible pointing, and marking by computer could be compared experimentally with handwriting and human grading of the same item types done with paper and pencil. These studies should compare score distributions, error rate, cost, and testing time under each condition.

Voice recognition technology might be used in the future to accept vocalizations as responses to CT test items. This opens up new opportunities for testing vocal utterances in language, or for examining preliterate or illiterate examinees, but it will, in all likelihood, also introduce measurement errors in identifying vocalized inputs.

Reduced measurement error. Computer-administered tests offer potentially significant reductions in several classes of measurement error. The elimination of answer sheets not only increases the speed of test taking, it might also eliminate traditional errors, such as penciling in the answer to the wrong item number, failing to completely erase an answer, and inadvertently skipping an item in the test booklet but not on the answer sheet. A further reduction in measurement error occurs with computerized tests because examinees can focus on one item at a time without being distracted, confused, or intimidated by the numerous items per page for paper tests. Computerized tests might therefore provide more accurate measures of performance for students who have lower reading ability, lower attention span, and higher distractibility.

Computer-administered tests can also reduce measurement error associated with the testing process. As the computer accepts responses from a keyboard, keypad, mouse, or touch screen, these responses are already in digitized form. They do not require a separate optical scanning step to put them in a machine-readable format. This provides opportunities to code the digitized responses with great sophistication, reducing the requirement for manual coding of short constructed answers or marks on a drawing. In addition, changes to answer keys, norm tables, and test-scoring algorithms can be more easily made with computerized testing than with paper-and-pencil scoring booklets, which require that every printed copy be updated or replaced. Computerized testing can also eliminate problems arising from lost or misplaced answer sheets and booklets, failure to mechanically scan pen and ink or faint pencil marks correctly, test scanner registration and resolution problems, and use of the wrong answer keys for scoring.

Computerized testing might, however, introduce new classes of measurement error. The need for multiple computer screens to read lengthy comprehension items could introduce new measurement errors. Multiple screens might introduce a memory component into the construct being measured. The use of shorter paragraphs to reduce the need for multiple screens has been shown by Green (1988) to affect the construct measured.

If the graphics resolution of the computer screen is not sufficient to produce quality equivalent to the graphics displays on paper-and-pencil tests, discrimination errors could interfere with measurement. Use of the response entry device, whether keyboard, touch screen, voice, or mouse, could also introduce new measurement errors. Additional research is needed to evaluate the effect of these new sources of measurement error. It may be that any new sources of measurement error attributable to CT will be dependent on item format, such as long reading passages, which might introduce subtle changes in the cognitive processes and, thus, the construct measured.

Ability to measure response latencies for items and components. The administration of tests by computer allows direct and precise measurement of response latencies. Latencies can be measured separately for specific item components (e.g. reading time for the item stem; analysis time for any complex drawing, graph, or table; reading time for each option; response selection time, or response speed). Dillon and Stevenson-Hicks (1981) and Sternberg (1982) note one consistent difference between good and poor problem solvers: the amount of time spent in mentally encoding reasoning problems. The good problem solvers spend more time reading and understanding the problem and the problem elements, whereas poor problem solvers quickly begin to try out solutions. Response latencies to individual test items, subtests, and total tests can be easily measured and collected. Precise measurement of any of these latencies is virtually impossible with paper-and-pencil tests. New psychometric models are needed to deal with response latencies. Although investigators have collected latency data, standards for using it to make valid inferences are rare. A fairly recent book by Luce (1986) is an exception.

Scoring and Reporting

Benefits to scoring. The process of scoring requires not only an accurate coding of each response, described earlier, but also combining individual item scores into meaningful subscores. When applicable, scoring also involves the computation of composite scores. The time required and the errors associated with applying complex subscoring procedures can be eliminated entirely, because the computer can calculate subscores and composites in the blink of an eye. Because complex subscoring procedures, such as those found in personality or biographical inventories, introduce so many possibilities for error, it is hard to disseminate these traditionally administered instruments without careful training and certification of examiners. Although computerized versions of personality or biographical inventories will still require examiner training to assure proper test use and interpretation, the training associated with scoring can disappear.

Benefits to reporting and interpretation. Computerized testing provides for immediate reporting of scores and offers many aids to human interpretation of the scores. Within a few minutes after completing the test, the examinee or the test administrator can receive a score report and prescriptive profile. Most standardized paper-and-pencil tests currently used require a minimum of between 6 and 9 weeks for test scoring and reporting. Many standardized tests have been criticized as having little direct instructional value because of these lengthy delays.

A specific example of the benefit of immediate scoring and reporting is provided by a recent large scale implementation of computer-administered testing by the Waterford Testing Center and at 39 schools in the Garland, Texas, school district (Slawson, 1986). The Waterford Testing Center developed a computer-administered version of the Garland PREDICTS test, a criterion-referenced and diagnostic test of reading, mathematics, and language arts at grades 3, 5, and 8. Using 34 WICAT 30-station supermicrocomputers, all third-, fifth-, and eighth-grade students in the district were tested within a three-day period. Immediately following testing, score reports and diagnostic and prescriptive profiles were prepared at each school for each of the individual students, classes, grades, and schools. The prescriptions included specific computerized lessons and textbook pages for the students to study. The score reports and diagnostic prescriptions were provided to all teachers within two days of test administration.

Although the test involved is part of the CT generation, the application goes a long way toward the use of measurement for individual purposes. If tests like PREDICTS were given frequently during the year in school districts, they would provide many of the benefits of third-generation testing. This example illustrates the power of the delivery system in advancing the generations of computerized testing. It would be unthinkable to invest the amount of money that the Garland District has for testing alone. The testing application came as a fringe benefit of the installation of a computer-aided education system with a substantial body of curricular materials.

Obtaining Records at a Central Site

When receiving and encoding responses, computerized testing produces a digitized version of the response vector, including latencies, if desired. This digitized record precludes the need for physical transportation, processing, and storage of voluminous paper bundles. Digitized data can be transmitted with a very low error rate to a central site where the data can be processed for item statistics, further analysis for research or educational decision making, and archival purposes. Transmission can be accomplished over telecommunication networks, or by mailing a magnetic disk or tape. The advantages of these processes over the collection, mailing, scanning, storing, and archiving of printed forms are obvious. Digital transmission and storage of records, however, complicates the administration of standards for fairness and, sometimes, state laws, which depend on access to printed documents with signatures to adjudicate disputes.

Automating Individually-Administered Tests

Individually-administered tests, such as the Wechsler Intelligence Tests, the Stanford-Binet Intelligence Test, the Kaufman Assessment Battery for Children, and the Luria/Nebraska Neuropsychological Battery, require standardized one-on-one administration by a trained examiner. Such tests will still require one-on-one administration by a trained person, but standardization and speed of administration could be improved but the administrators of these computerized tests would not require as much training. Interpretation and proper test use should continue to be under the direction of trained professionals.

To achieve these goals through CT, the computer may be programmed to interact primarily with the test administrator. It can prompt the administrator about which objects to arrange for performance tests, and it can provide easy response formats for entering the coded result for each item involving the interpretation of vocal responses or movements. In addition to prompting the examiner for items requiring interpretive judgment, many of the items and item types in these tests may be presented on the computer display under the precise controls described earlier. In these cases, the examinee could respond by pointing, by pressing a key, or, in the case of older students, by typing a few words. A touch- sensitive screen, a mouse, or cursor arrow keys allow students to point to responses.

CREATING TESTS AND ITEMS BY COMPUTER

This chapter has deliberately been narrowed to focus on the processes of test administration, excluding computer applications to development and to analysis and research. It is impossible not to mention some implications for these other areas, however. Collection of data at a central site for calibration and computing of item statistics, for example, is closely related to test development.

Computer-aided Test Assembly

Closely associated with the first generation is the wider dissemination of computer aids to test assembly, using item banks and tools involving word, text, and graphics processing to aid in the process of creating items. Electronic publishing and fast laser printers make localized, or even individualized, paper test forms feasible in some applications. Products are now being introduced to permit users to create customized tests and items measuring individual goals and objectives of schools, districts, educational service agencies, and state departments. With such software for test creation, educational agencies are able to select the grades, subjects, and objectives of the tests to be created, review the domain specifications or expanded objectives, select the specific items to be included in the test, sequence the objectives and items, and create an operational test to be administered by computer. Paper-and-pencil tests can also be created with the software and printed out on a laser printer. For computerized tests, the resulting software often includes all necessary modules for test registration, scheduling, management, administration, scoring, reporting, and providing specific curriculum prescriptions.

These applications are discussed in Olsen et al. (1984) and Slawson, Maynes, Olsen and Foster (1986).

Computer-created Tests

Instead of storing an item bank with fixed item contents and formats, the computer can also be used to create tests and items using a bank of item generation algorithms and item forms (see Millman, 1977, 1984a, 1984b; Baker, in press). Such a bank would contain several hundred item skeleton structures or item forms. Through a series of interactive screen displays, developers can specify item-content elements, item formats, and scoring options that can be used by the computer to generate approximately equivalent items from the same content domain.

RESEARCH ISSUES FOR THE CT GENERATION

The fundamental research question for the first generation is the equivalence of scores between a computerized version of a test and the original, paper-and-pencil version. This is also an important issue for the second generation. For some time, testing organizations will wish to give users a choice between a paper-and-pencil or other conventionally administered version and a computerized version. The question of score equivalence will thus be fundamental for some years, as computerized tests become more widespread. Few of the underlying differences between CT and conventional tests are of lasting scientific interest, so as score-equivalence studies are completed, research on CT will shift to scientific issues dealing with matters like individual and group differences, how to use latency information, and what constructs are measured by advanced forms of computerized tests.

The American Psychological Association Committee on Professional Standards, together with the Committee on Psychological Tests and Assessment, developed a new set of APA Guidelines for Computer Based Tests and Interpretations (1986). They outline the conditions under which scores can be considered equivalent. The rank orders of scores of individuals tested in alternate modes must closely approximate each other. The means, dispersions, and shapes of the score distributions must be approximately the same, either directly or after rescaling the CT scores. The Guidelines hold test developers responsible for providing evidence of score equivalence. This might be an expensive proposition for test developers, requiring separate equating and norming studies for CT versions of tests, at least in those circumstances wherein the scores from the two modes of administration are to be used interchangeably. Costly equating and norming studies are a barrier to the introduction of computerized tests. Research that might show the circumstances and design features of computerized tests under which equivalence could be assumed would therefore be beneficial in advancing the field.

The current pool of research in this area is quite shallow. The most recent review, commissioned by the College Board and ETS, was conducted by Mazzeo and Harvey (1988). This review identified fewer than 40 studies comparing computerized and conventional tests. A number of the earlier studies did not consider computerized testing as we know it today. Today, we assume a cathode-ray tube or some other kind of electronic display and a keyboard or pointing device for response entry, but several of the studies reviewed presented test items on projected colored slides and used a variety of response mechanisms, including paper and pencil.

Table 2 summarizes the results of a representative set of the studies reviewed by Mazzeo and Harvey (1988). In general, it was found more frequently that the mean scores were not equivalent than that they were equivalent; that is, the scores on tests administered on paper were more often higher than on computer-administered tests. The score differences were generally quite small and of little practical significance. A major exception to this was the Coding Skills Tests. These are speed tests in which the speed of responding on the computer keyboard greatly favors the computer group. Scores are also affected in certain computerized personality tests, where omit rates were quite significantly higher on computer-administered tests than on paper-and-pencil tests (indicating "cannot say" rather than "true" or "false" in response to a personality statement). Mazzeo and Harvey expressed concern that differential omit rates might also occur on other kinds of tests, affecting formula scoring on ability tests and personality scores. The personality subscores of the MMPI were reduced by the higher frequency of choosing "cannot say" and perhaps by other factors.

The superior performance on paper-and-pencil tests may be due to artifacts; for example, the study by Hedl, O'Neil, and Hanson (1971), showed a large mean difference in favor of the paper-and-pencil group on the Slossen Intelligence Test. The computer scored the typed responses automatically and could have introduced scoring errors in this

Table 2

Research Studies Contrasting Score Equivalence of Paper-and-pencil and Computerized Tests

Type of test	CT scores higher than paper-administered	CT scores lower than paper-administered	No significant differences
Free-response tests		Elwood, 1972a; 1972b; Hedl et al., 1971	
Computerized personality tests		Biskin & Kolotkin, 1977; Lushene, O'Neil, & Dunn 1974; Scissons, 1976	Lukin, Dowd, Plake & Kraft, 1985; Parks, Mead & Johnson, 1985; White, Clements, & Fowler, 1985
Aptitude Tests	Johnson & Mihal, 1973 (for Black examinees)	Lee & Hopkins, 1985; Lee et al., 1986; Sachar & Fletcher, 1978 (timed test)	Johnson & Mihal, 1973 (for White examinees); McBride & Weiss, 1974
Achievement tests			Olsen et al., 1986; Wise, Boettcher, et al., 1987; Wise & Wise, 1986
Coding skills tests	Greaud & Green, 1986; Kiely et al., 1986 (one item per screen)	Kiely et al., 1986 (numerical items)	
Graphics tests		Jacobs, Byrd & High 1985; Jonassen, 1986	Reckase, Carlson & Ackerman, 1986 (untimed); Kiely et al., 1986
Multiple page items		Kiely et al., 1986	Feurzeig & Jones, 1970

study. Only small differences were found by Elwood (1972a, 1972b), who had the examinees type in the answers to questions on the Weschler Adult Intelligence Scale, but scored the responses by hand.

Neither Mazzeo and Harvey (1988), nor the current authors are willing to make the generalization that computerized testing is more likely, in general, to lead to slightly lower scores. Indeed, as Table 2 shows, the reverse is often true. Another reason to doubt this generalization is that most of the studies reviewed suffered from several kinds of confounding difficulties, which will be discussed in the sections below on engineering and experimental design issues.

ENGINEERING DESIGN ISSUES

The field of computerized testing has not yet matured to the point where consistent specifications exist for the interface between testee and material for each item type. Consistent design standards are needed to administer items, to provide access to different parts of an item that requires more than one screen to display, and to provide a way to immediately correct response entry errors or to change an earlier item. The last process can be accomplished with paper-and-pencil tests by erasing the marks on the paper answer sheet. The lack of consistent interface engineering standards was suspected by Mazzeo and Harvey (1988) and, in many cases, by the original authors cited in Table 2, as a causative factor in the score differences. For example, in the study by Lee et al. (1986), 585 naval recruits took a paper-and-pencil version of the ASVAB Arithmetical Reasoning Test between two and six months before taking a computerized version. Questions were presented one at a time on a computer terminal, but subjects could not refer to previous items or change answers. The mean number-right score in the paper-and-pencil condition was about one point higher (on a 30-item test) than in the CT condition. Unfortunately, the interface design did not permit subjects to correct immediate key entry errors made in entering the responses or to review and change previous items. Of course, this was possible in the paper-and-pencil version of the test. This is a trivial engineering design problem for which several good solutions exist.

Engineering and language-processing research are needed to improve the accuracy of identifying computerized free-response answers. Some possible variables are spelling tolerance algorithms, synonym dictionaries, and ignorable word lists. In the fourth generation, artificial-intelligence capabilities for processing natural language will become available. These capabilities might have a substantial impact on this issue.

Another engineering design problem lies in providing simple and effective conventions for reviewing previous parts of a large textual item (e.g., a paragraph comprehension item) or a text-plus-graphics item. These standards are needed when the entire item cannot be displayed simultaneously on one screen. One good design solution is to make the question visible in a foreground window while the previous material is paged or scrolled rapidly in a background window. Kiely, Zara, and Weiss (1986) recommended a variant of this solution in connection with a study of paragraph comprehension items. Each of three different CT conventions for reviewing parts of the paragraph, with and without the question visible, produced lower mean scores for the CT students than for students taking the same items with paper and pencil.

BEST COPY AVAILABLE

Keeping the question and related parts of the paragraph visible simultaneously reduced the advantage of the paper-and-pencil group.

Paging among multi-screen items, whether text or graphics, is an engineering design problem that interacts with the resolution of the screen and with the windowing or scrolling conventions adopted. The trend in the field of user interfaces is more and more toward higher resolution screens, some equivalent to a page of text. The trend is also toward the scrolling and windowing conventions available on the Apple Macintosh computer and available on IBM and compatible equipment through the Microsoft Windows package. The speed and effectiveness with which a test taker can use these conventions interacts not only with the resolution of the display and the flexibility of the software, but also with the user's familiarity and facility with the conventions. Familiarity with conventions creates an experimental design issue, which was also discussed in the Mazzeo and Harvey review.

EXPERIMENTAL DESIGN ISSUES

Mazzeo and Harvey (1988) found that the effects of practice were significantly different on tests taken in the two different modes. These practice effects in some cases confounded the results and in other cases produced puzzling interactions. In several studies, Mazzeo and Harvey reported that the increase in scores was likely to be larger when the automated test was administered before the conventional test. As a result of these asymmetric practice effects, the authors argued against conducting equating studies based on single-group counterbalanced designs. The authors also expressed frustration in interpreting the results of a number of equating studies, which perfectly confounded alternate forms of the test with computer versus conventional administration or confounded the order of administration, intervening learning, and other factors.

User familiarity with a computerized testing interface is an important consideration in conducting equating studies between paper-and-pencil and computer administration. As the results in Figure 1 suggest, taking a test with a mark-sense answer sheet and a booklet is a learned skill. Third graders showed considerably less facility with it than sixth graders. Similarly, facility with a particular set of computer interface conventions for moving among pages of an item, changing answers, and reviewing parts of current and previous items is also a learned skill. Most of the studies reported by Mazzeo and Harvey did not provide assurance through instruction and practice that the examinees were familiar with and had facility with the particular interface conventions. The study by Olsen et al. (1986) provides an important contrast. In this study, the third and sixth graders had already had considerable time to familiarize themselves with the computer and its interface conventions. The 30-terminal computer system had been installed in their school long enough prior to the study that use of the computer for instruction and testlike items had become a familiar routine. In this study, score equivalence was found. Nevertheless, a considerable reduction in testing time was shown for computerized testing. The tests they studied were nonspeeded power tests.

The differential effects of computer administration versus conventional administration on the speed of test taking have now been demonstrated in several studies besides Olsen et al. (1986). Greaud and Green (1986) and Kiely et al. (1986) found it affected score differences on coding skills tests. Kiely et al. (1986) also found, however, a smaller effect in favor of paper and pencil in the numerical operations speeded test. Sachar and Fletcher (1978) found that the engineering design of a feature, or perhaps a particular computer's inherent speed in reviewing and correcting previous items, could have slowed the computer group down sufficiently that they completed fewer items in a speeded aptitude test, thus reducing the scores for the CT group. Speed effects are apt to differ between CT and conventional tests. When tests are speeded, there appears to be a three-way confounding of engineering design (including the speed of a particular computer in retrieving and displaying previous items), familiarity with the interface, and differential speed limits of human responding inherent to each medium. Anyone who has observed the response speed learning curve for a teenager on a complex computer game is well aware of the incredible levels of psychomotor speed that can be obtained through practice on a computer interface. It is doubtful that responding with a pencil on a mark-sense page has as high an upper limit for speed and simultaneously retains an acceptably low error rate.

The problem of equating computer and conventional speeded tests might not be an easy one to solve and could ultimately require separate norms. On the one hand, numerous studies show a speed advantage for CT. On the other hand, Sachar and Fletcher (1978) indicate that a likely cause of effects from mode of administration with the speeded power tests of aptitude was the amount of time associated with the error correction and review features of the computer version. It is impossible to judge from these studies what the differential effects would be, given well-designed, well-practiced interface conventions. Some item types (e.g., paragraph comprehension with low-resolution screens) are likely to be fundamentally slower with the computer, some faster (e.g., coding speed items).

SCIENTIFIC ISSUES

As engineering and experimental design issues become better stabilized and the methodologies for equating conventional and computerized testing become familiar and standardized, emphasis will likely shift toward scientific issues of greater import. Chief among these is the construct validity of new computerized measures. After all, mean differences disappear after equating; hence, they mean nothing so long as the two tests measure the same construct. Measures that are difficult or impossible to obtain by conventional means will be especially interesting subjects for construct validation. Items involving animation, motion, and audio presented by the computer need to be investigated. What constructs are measured? Are they new or the same constructs measured by conventional methods? What, in particular, are the meanings of response latency, presentation time, and other aspects of temporal control in terms of the constructs measured?

Individual and group differences. It could be that individual or group differences affect examinee performance on conventional tests versus CT. For example, research by Wise, Boettcher, Harvey, and Plake (1987) has shown nonsignificant effects of either computer anxiety or computer experience on conventional versus CT versions of the same test.

D. F. Johnson and Mihal (1983), however, compared the performance of Black and White examinees on the paper-and-pencil version, versus a CT version, of the Cooperative School and College Ability Test. The average total test score of the Black students was 5.2 points higher (on the 100-item test) on CT than on conventional testing. The White students only scored .5 points better on total score with CT. Johnson and Mihal's interpretation is that the Black students were more motivated, due to the novelty of the CT environment. Futhermore, these authors state, their scores were less likely to be depressed by the negative expectations those Black students might have had toward intelligence and aptitude measurement.

Wide differences in well-practiced response speed might prove to be an important variable in individual differences. Response speed varies with age, and the sexes might be differently motivated to practice more or practice less. This individual difference in response speed could affect scores on CT tests involving pure speed and speeded power or when paging or keying facility is required.

Grading of Performance and Verbal Productions with Human Assistance

Research is needed on using computers to aid paraprofessionals in administering and scoring individualized tests requiring observation and judgment of performance tasks and verbal productions. This is a new frontier. Little research has been reported on the adaptation of expensive, individually-administered tests to computerized format to aid a human test administrator or observer who responds to the student when judging vocalizations and movements. This approach would broaden the opportunity for administering such tests. It would also provide a new alternative for testing preliterate, illiterate, or handicapped individuals who cannot read instructions from the screen, or who are unable to respond by using conventional computer input devices.

THE SECOND GENERATION: COMPUTERIZED ADAPTIVE TESTING (CAT)

The primary difference between the CAT and the CT generations is that CAT tests are administered adaptively. This, of necessity, generally requires greater computer speed and computational capability and advances in psychometrics, including generally more sophisticated measurement scales. Adaptive tests provide even greater speed of administration than CT, because fewer items need to be administered for equal or greater precision.

DEFINITION

The second, or CAT, generation of computerized educational measurement is defined as computer-administered tests in which the presentation of the next task, or the decision to stop, is adaptive. A task can be an item or a more complex standardized situation involving one or more responses. To be adaptive means that the presentation of the next task depends upon calculations based on the test taker's performance on previous tasks.

The calculations required to select the next task might require additional computer capabilities, such as floating-point arithmetic, and a faster processor than is required for minimal first-generation tests. Item response theory (IRT) provides a psychometric foundation for one kind of CAT test, that which adapts primarily on the basis of the item-difficulty parameter. This type of test measures static position on an interval scale and has initially been used for primarily institutional purposes, such as selection or placement. CAT tests may also be used for individual purposes.

EXAMPLES OF COMPUTERIZED ADAPTIVE TESTS

Three cases of adaptive tests will be described and an example given of each: adapting item presentation, based on IRT parameters, particularly the difficulty parameter; adapting item presentation times, based on previous response times; and adapting the content or composition of the item, based on previous choices. In any of these cases, a separate adaptive decision may be made: adapting test length, based on the consistency of previous performance.

Adapting item presentation based on the basis of IRT parameters is the best understood among possible adaptive tests. It will be discussed in some detail later in this section. The other two types of adaptive tests will be discussed first.

Adapting Item Presentation Speed

A computer-administered test adaptive on item presentation speed was developed by the authors at the WICAT Education Institute in 1983 as part of the first experimental learner profile battery. The test was designed to assess a construct of perceptual speed. This construct involved processing that was presumed to require both cerebral hemispheres. Called the Word/Shape Matching Test, it contained items involving both a word and a shape that either matched or did not match. An example stimulus would be the word circle with a square drawn above it. The examinee was instructed to strike one key for a match, another for a mismatch. If a choice was not made within a certain time, the item would time out, and another item would be presented. In this case, more time would be given for the next item. Correct responses within the current time interval would lead to a shorter interval the next time. Two scores were provided, percentage correct and the asymptotic time interval. It was hypothesized that the speed score would be a good predictor of success in other speeded tasks requiring processing of both pictorial and verbal stimuli; that the willingness to trade off speed for errors would be a useful indicator of cognitive style; and that low scores would be one indicator of potential learning disability. Confirming or disconfirming these hypotheses required a long-term research program that we were not able to complete. The test was administered to groups of elementary school students, who had no difficulty with the operation of the test.

Adapting Item Content

Simulation tasks are always adaptive because the next piece of content to be presented depends on the responses of the user. Some simulations also adapt on the basis of endogenous events, like the passage of a certain amount of time. Simulation items could be the most sophisticated contribution of computerized measurement to increased complexity and interest in testing. These often use a computer-controlled videodisc to present simulated displays adaptively. Examples are a patient having a medical examination, a piece of equipment needing repair, and images of fruit flies in a genetics breeding experiment. The student makes a series of decisions in the simulated environment, and scoring is accomplished by evaluating the outcome, strategies, and sophistication of the path followed. The National Board of Medical Examiners has investigated patient-management simulations extensively as a possible part of their medical certification examinations. They organized a short-lived company, Computer-based Testing and Learning, Inc., to develop and administer these new tests (National Board of Medical Examiners, 1987). The face validity and user acceptance of these tests is high, but the industry is immature.

Simulations are not the only example of the adaptation of content on the basis of examinees' responses. A computer-administered test that adapted the next paired comparisons was developed by the authors in 1983, as part of the initial WICAT Learner Profile Battery. This test produced a preference profile on the two bipolar dimensions of analytical and logical thinking versus feeling and interpersonal preference and of intuitive, holistic processing versus controlled, sequential processes and preferences. Items were forced-choice, paired comparisons involving an illustration and a phrase such as "I like to hug," "I like to take things apart to see how they work," "I like to draw pictures of imaginary things," "I like to keep my desk neat and tidy." As in a tournament, winners were paired with winners and losers with losers, until a complete ranking of most preferred to most avoided statement was obtained.

Most simulation tests, including the Learner Profile tests described, should be considered experimental. Considerable work is needed to establish strong psychometric foundations for tests adaptive on presentation time, content presentation, or any other basis. With tournament-style ranking, for example, single elimination does not provide enough data to assess or to assure reliability of the ranking.

ADAPTING ON THE BASIS OF ITEM PARAMETERS

Adapting the order of item presentation on the basis of IRT parameters is based on several decades of psychometric research. In the remainder of this section, the term adaptive test will refer to tests adaptive on IRT parameters.

In a conventional test administered by either paper and pencil or computer, the majority of items are too easy or too hard for a given examinee and the examinee will likely answer all easy items correctly and miss the more difficult ones. The items that are too difficult or too easy will contribute little information for measurement of the person's true ability level.

Although the basic ideas and methods of adaptive testing have extensive historical roots in the work of Binet (1909), Birnbaum (1968), and Lord (1970), it was not until the development of digital computers that adaptive testing became feasible. The computer can quickly calculate ability and error estimates and check to see if the criterion for test termination has been met. With a computerized adaptive test, each examinee can be measured at the same level of accuracy or precision. In contrast, with conventional paper-and-pencil or CT tests, the scores near the mean are measured more accurately than those at the high or low end of the score scale.

A computerized adaptive test requires a large item bank that has been calibrated in advance to yield parameters fitting a theoretical item response curve. Each theoretical curve is a function relating probability of correct response to the underlying latent-trait dimension. The computerized adaptive test also requires that responses to items be locally independent (not influenced by responses to any other items). A further assumption is that the items in the bank for a subtest measure a single underlying unidimensional ability or latent trait. Psychometric research is currently underway to develop multidimensional models for use with computerized adaptive testing (Reckase, 1985; Hambleton, in press).

Steps in Administering an Adaptive Test

There are four major steps in administering an adaptive test.

1. A preliminary estimate of ability is made for the examinee.
2. A test item is selected and administered that will provide maximum information at the estimated ability level. The information value of the item can be calculated on line or stored in a precomputed information matrix. Generally, if the examinee answers an item correctly, a more difficult item is presented; if the examinee misses the item, an easier item is administered. Of all the items available, the one selected is calculated to maximize new information about that examinee, subject to constraints due to content balance and limits placed to control excessive exposure of certain items.
3. The ability estimate is updated or revised after each item. A variety of methods have been proposed for ability estimate updating (Hambleton (in press)). The methods proposed include Bayesian Sequential Ability Estimation (Owen, 1969, 1975), Maximum Likelihood Ability Estimation (Birnbaum, 1968; Lord

1977, 1980; Samejima, 1977), Expected A Posteriori Algorithm (Bock & Aitkin, 1981; Bock & Mislevy, 1982a) and biweighted Bayes estimates. The biweighted Bayes is a robustified ability estimator (Bock & Mislevy, 1982c; Jones, 1982; Wainer & Thissen, 1987, Wainer & Wright, 1980).

4. The testing process continues until a designated test termination criterion has been met. Typical termination criteria include a fixed number of test items, when the standard error reaches or is less than a specified value, and when the test information function reaches or exceeds a specified value.

Item Response Theory

Computerized adaptive testing is based on the psychometric theory called item response theory (IRT) developed and explicated by Birnbaum (1968), Hambleton (in press), Hambleton and Swaminathan (1984), Hulin, Drasgow, and Parsons (1984), Lord (1952, 1970, 1980), Lord and Novick (1968), Rasch (1960), and others. IRT postulates that examinees differ in their ability on a unidimensional continuum ranging from low to high ability. For each examinee, the probability of answering each item correctly is dependent on the current ability estimate of the examinee and the properties of the item response curve for the current item. Item response curves are usually specified by up to three parameters, the location of their most effective point (the difficulty), their slope at that point (the discrimination), and their *y* intercept (guessing parameter).

Calibrated Item Banks

CAT tests require the careful development and calibration of a relatively large pool of items. The usual minimum number of items in a pool is 100. These items are administered to a large number of examinees from the target population, and response vectors are obtained for each examinee. With the data from five hundred to one thousand response vectors, calibration programs are used to estimate the parameters of the chosen item response curve. Once calibrated, items can be added to the operational item banks and used in CAT systems. A program for continual update of item banks by obtaining new response vectors for calibrating new experimental items is generally part of an operational CAT system. New experimental items are introduced into the item banks and are administered on a prescribed schedule. These items are not part of CAT scoring, but they are part of the process of developing a sufficient number of response vectors to calibrate the new items.

The study of the difficulty parameter, the discrimination parameter, and the guessing parameter represents a powerful form of item analysis and can be used to refine or discard experimental items. As an item analysis technique, a generalized form (Thissen & Steinberg, 1984) can even provide information on the attractiveness of distractors, but IRT item analyses have usually been considered deficient in this respect.

Several alternative item calibration programs have been developed. The most widely used are LOGIST (Wingersky, Lord, & Barton, 1982) and BILOG (Mislevy & Bock, 1982). More recent candidates include ASCAL

(Vale & Gialluca, 1985), MICROSCALE (Linacre & Wright, 1984), M-SCALE (Wright, Rossner, & Congdon, 1984), and MULTILOG (Thissen, 1986).

CURRENT COMPUTERIZED ADAPTIVE TESTING SYSTEMS

With the emergence of microcomputers and low-cost multiprocessors, computerized adaptive testing has now become feasible for widespread operational research and implementation. Within the past few years, a variety of microcomputer-based adaptive testing systems have been developed, demonstrated, and implemented. The military has sponsored the most far-reaching and complex development projects.

The first of the military CAT system prototypes was developed for the Apple III computer by the Naval Personnel Research and Development Center (Quan, Park, Sandahl, & Wolfe, 1984). This prototype was developed to provide computerized adaptive administration of the subtests from the Armed Services Vocational Aptitude Battery (ASVAB). Following successful research on the validity and reliability of the computerized adaptive ASVAB, compared with the paper-and-pencil ASVAB, the Department of Defense contracted with three independent companies to design and develop operational CAT systems (WICAT Systems; Bolt, Beranek and Newman; and McDonnell Douglas). The military elected not to complete the procurement process initiated with these contracts, but important lessons were learned for large-scale CAT development.

These system prototypes were developed for future administration of a computerized adaptive ASVAB in 69 military enlistment processing stations in larger cities and in up to 800 smaller mobile examining team sites. This arrangement requires a large fixed-site configuration, a small portable configuration, and the communications to link them to one another and to a specified military base for central record keeping.

The components of an operational CAT system are numerous, involving complex hardware and software. One of these prototype CAT systems included a portable supermicrocomputer system (the WICAT 150) with the powerful Motorola 68000 CPU chip. This processor had a multiprocessing operating system and the speed to handle up to eight portable graphics terminals simultaneously. The graphics resolution was sufficient to display effectively the line drawings found in ASVAB items, such as mechanical comprehension and automotive information items. The system developed by Bolt, Beranek, and Newman also supported eight graphics terminals from one central portable processor, but the McDonnell Douglas prototype used a separate CPU for each display.

The military procurement process is extremely thorough. It applies standards for the development of hardware, for software and applications programs, and for human factors that are not always considered by a civilian user who is contemplating system procurement for the transfer of widely distributed paper-and-pencil testing programs to computer. The military configuration is applicable to many civilian organizations that administer tests. These organizations frequently require fixed locations in major cities and portable systems that can administer tests on a less frequent basis to smaller groups of people in temporary locations. They often have one central site at which the test scores are archived and personnel decisions are made. A number of professional or membership organizations who test for admission or certification of individuals for practice in a profession have similar requirements. For these reasons, we will summarize some of the features such a large-scale user might look for.

Selection of Hardware

An organization deploying a large testing configuration would profit from the use of standard hardware components, including computers, buses, peripherals, and interface devices. Availability of a maintenance network and a strong record of reliability should be strong considerations. Features to enhance maintainability include power-on testing, device initialization, failure logging, diagnostic monitors in read-only memory (ROM), and diagnostic downloading from a host to a remote computer.

Other maintenance considerations include modularity of parts, ease in connecting and disconnecting parts and cables, and uncomplicated cabling. The user should also be concerned with safety and electronic emission standards.

Human factors. General operating factors, such as table arrangements, good lighting, avoidance of glare from windows, and electrical requirements are important in all circumstances, but they must be considered each time for mobile operations in a temporary site. The portability, size, and weight of the equipment are of course the major factors in such operations. The attractiveness for possible pilfering of keypads and other small components must be considered in their design. They should not be easy to detach.

Legibility and visibility factors of the screen include the luminance, contrast, resolution, display size, effective viewing angle, viewing distance and glare, as well as jitter and drift of the screen.

Proctor control of video adjustment controls is desirable. These are factors inherent in the hardware. Also important are what has earlier in this chapter been called the interface conventions. Good engineering design provides unambiguous screen formats, including a standard format for each item type; a clear set of conventions for paging or scrolling on multi-page items; the visibility of the question in the foreground window on multi-page items; a rapid response to the examinees' inputs so that response speed is not affected by computer delays; a legible type font; and clear and legible graphics.

The keyboard and keypad input factors include the ability to time responses and the ease with which the keyboard and its conventional uses with different item types can be learned and used. Also important is a compact size, an auditory click or sound to indicate engagement of a key, and the "feel" of the keys. If touch screens are to be used, parallax and the resolution of the active touch sites are important to consider. A joy stick, a mouse, or curser-control arrows introduce differences in speed of response, in sources of error, and in the learning curve for fast and accurate responses.

A printer will always be associated with the computer at a larger fixed location, and a portable printer might be available for mobile applications. In this case, legibility, standard format and use, and speed of issuing the reports after testing are important factors.

In connection with the user friendliness of the system, not only must the user interface conventions be quick and easy to learn, but familiarization should be provided at the beginning of the entire testing session and at the introduction of each new item format to assure that users are familiar with the conventions. Computer literacy and standard input and display conventions are rapidly becoming more widespread in our society. Any CAT system should use familiar and widely accepted conventions and should not introduce unusual conventions that require the user to change established habits.

Selection of Software

Software subsystems can be grouped in three categories; software for the central development facility, software for the fixed sites in major cities, and software for operations at both the fixed site and the portable systems.

Software at the central development site. The authoring software should permit the simulation of a test for tryout by the developers, with full debugging tools. Item authoring programs should permit flexible screen editing of text and graphics. The ability to enter items in a batch mode from existing text files is a significant advantage that enhances productivity. Editing features should include access to files to update the test battery composition and composite score weights. Speeded (timed) test and item authoring might require a separate software module, as could familiarization sequence authoring. Item calibration programs must be installed at the central development site, generally on larger computers. Programs to insert experimental items into test batteries and to schedule their introduction must be available. Encryption and decryption software must be available if items, item banks, and other secure information are to be transmitted to the fixed sites electronically or on a magnetic medium that might be intercepted. Communication software will be extensive at the central site, as it communicates with all fixed sites.

Software at the fixed locations. Item data collection and data consolidation programs are necessary at the sites to consolidate the data collected locally and at the satellite mobile sites reporting to a given fixed site. A test-score and student record archiving function must also be available. Communication software to transmit the archived scores to the central site where personnel records are kept is necessary, along with communication software to communicate with the portable computers at temporary sites. Finally, communications with the central test-development site are necessary. This site might be different from the central site for personnel records.

Software at the temporary and fixed locations. Programs are needed to register the testees and to obtain necessary biographical information. For military enlistment, this includes some medical information. Programs are necessary to assign each examinee to a particular workstation, to log the examinee in, and to provide computer familiarization. For the proctor or the operator with a portable computer, software should boot the computer, monitor the terminal ports from one proctor terminal, provide password security for the proctor versus the test taker, and permit control of the printer. Software to administer the test includes a program to provide the initial estimate of ability and instructional programs for familiarization with each subtest, including practice with the user interface conventions. Also needed are item-selection programs, programs to accept and code examinees' responses and response times, ability estimation routines, and programs to check the test-termination criterion. Software must record the subtest results, provide a subtest report and a consolidated report, and file the test results by each experimental item and by each examinee, so that response vectors can be obtained for calibration. The portable computer must have communication software and a modem to communicate with the fixed site. It must also have encryption and decryption software, if this degree of security is required.

Various software routines are necessary to assist the operator at the fixed site to check the integrity of files, to control the communications, to manage the disks, to provide maintenance functions, to configure and reconfigure the system for different numbers of terminals and different numbers of computers, and so on. Proctor functions are also needed at both the central and mobile sites. In addition to logging examinees and assigning them to a particular terminal, the proctor needs to be able to monitor the terminals and watch for signals of trouble at any one (a raised hand is sufficient at the mobile site). The proctor needs software to stop a test and restart it at another terminal in the case of a breakdown midway through a test. Other programs are needed to delete old files and records, back up daily work onto disks or tapes, and restore records from a magnetic medium.

The amount of storage required for this extensive software is surprisingly low for at least one of the prototype systems. The data for test items for nine ASVAB subtests consisted of less than two megabytes. Each of the nine subtests included about 200 items with 10 or 20 graphics and required 220 kilobytes. The examinee data for a fixed site might be maintained at about two megabytes. The CAT software only required about 800 kilobytes, and the file system overhead needed about 200 kilobytes. This amounted to five megabytes, which easily fit on a ten megabyte hard disk. The procuring organization should consider much larger files. Hard-disk drives of much larger capacity are now available, and programs and data seem to obey a law that they always expand to fill the available space.

Other CAT Systems

Over the past few years, several CAT Systems have been developed by professional educational and psychological testing organizations. These systems are far less complicated than the military prototypes, except in the case of some test-development software subsystems. Current systems generally operate on personal computers. Educational Testing Service and the College Board have developed a CAT testing system for implementation on the IBM PC for measuring college level basic skills in English and mathematics (Abernathy, 1986; Ward et al. 1986). The Assessment Systems Corporation has developed a generalized microcomputer adaptive test-authoring and administration system (MicroCAT) implemented on an IBM PC (Assessment Systems, 1985). The MicroCAT system is now used by the Portland public schools and Montgomery County public schools for development of school-based adaptive testing. The Psychological Corporation has developed an adaptive version of the Differential Aptitude Test for administration on Apple II computers (Psychological Corporation, 1986). They have also demonstrated a computerized adaptive Mathematics Locator test for administration on the Apple II computer (McBride & Moe, 1986). The Waterford Testing Center has developed a generalized CAT authoring, administration, and reporting system. This CAT system has been used to develop a school-based learner profile aptitude battery consisting of 45 different CT and CAT tests and a comprehensive school- based computerized testing system for grades 3 - 8 that measures achievement in reading, mathematics, and language arts (WICAT Systems, 1988). The Waterford Testing Center has also developed CAT tests of mathematics applications for the California Assessment Program (Olsen, Maynes, Ho, & Slawson, 1986). Unlike the personal computer implementations, these systems operate on the 30-terminal WICAT Computer-aided Education System for larger fixed sites primarily engaged in instruction, but they are also available on smaller configurations.

ADVANTAGES OF COMPUTERIZED ADAPTIVE TESTS

Because computerized adaptive tests are also administered by computer, all of the advantages over paper-based testing noted for the CT generation also apply to CAT. In summary, these advantages are:

1. Enhanced control in presenting item displays
2. Improved test security
3. Enriched display capability
4. Equivalent scores with reduced testing time
5. Improved methods for obtaining and coding responses
6. Reduced measurement error
7. Ability to measure response latencies for items and components
8. Improved scoring and reporting
9. Automation of individually-administered tests
10. Obtaining records from a central site

11. Ability to construct tests and create items by computer

IRT also provides many advantages in the score equating process because each item has a calibrated position on an underlying latent-trait scale. Additional advantages of computerized adaptive tests are presented next (see also Green, 1983; Wainer, 1983, 1984; and Ward, 1986).

Increased Measurement Precision

Research has shown that conventional tests administered by computer or by paper have high measurement precision near the average test score but they have low measurement precision for low- and high-ability levels. In contrast, a computerized adaptive test maintains high measurement precision, or accuracy, at all ability levels (low, average and high). Setting the CAT test-termination criterion at a specified value allows all examinees to be measured to the same level of precision.

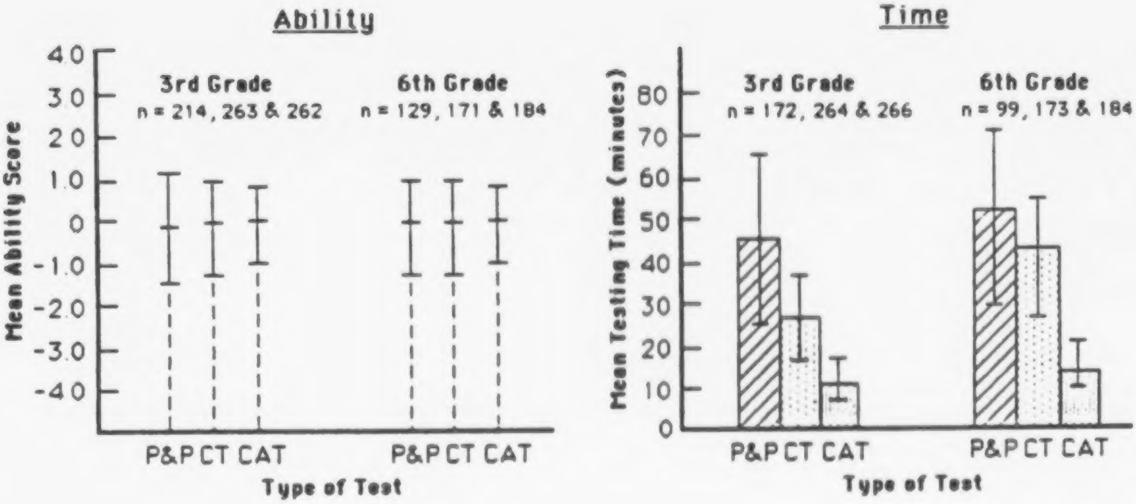
Equivalent Ability Estimates with Reduced Testing Time

With CAT, each examinee is administered only the subset of items similar to his or her ability level. Items too difficult or too easy for a given examinee are not administered. Compared with conventional paper or computerized tests, a CAT test requires far fewer items. The strongest claims for fewer items have been in the area of school-based achievement tests (Olsen, Maynes, Ho, et al., 1986). These authors found that only 30% to 50% of the test items were needed to reach an equivalent level of precision, as compared with paper testing. Correspondingly, less test administration time was needed. Ward (1984) notes that, with adaptive testing, "the length of a test battery can be cut by 50 to 60 percent and still maintain a measurement accuracy equivalent to that of the best standardized conventional test" (p.17). The research on score comparability of paper-and-pencil, computer-administered, and computerized adaptive tests is quite complex. As noted in the last section, computerized tests do not always yield equal ability estimates and distributions, but a case in which they do is presented by Olsen, Maynes, Ho et al. (1986). Figure 2 summarizes the key results. Equivalent ability estimates with reduced testing time have also been found by McKinley & Reckase (1980, 1984a), and Moreno, Wetzel, McBride & Weiss (1983).

Further Improvements in Test Security

Test security is enhanced with CAT beyond CT because each examinee receives an individually tailored test. It would be difficult to steal and memorize each of the hundred or so items for each of several such tests. Encryption can be used to protect item banks and examinee data.

FIGURE 2


Ability
Mean Ability Score
4.0
3.0
2.0
1.0
0
-1.0
-2.0
-3.0
-4.0
3rd Grade
n = 214, 263 & 262
6th Grade
n = 129, 171 & 184
P&P CT CAT
P&P CT CAT
Type of Test
Time
Mean Testing Time (minutes)
80
70
60
50
40
30
20
10
0
3rd Grade
n = 172, 264 & 266
6th Grade
n = 99, 173 & 184
P&P CT CAT
P&P CT CAT
Type of Test

Randomization can be implemented to select one of a set of most informative items. With CAT, there are no paper copies of the tests, answer booklets, or answer keys. Two examinees sitting next to one another are even less likely than with CT to see the same item at the same time.

RESEARCH PROBLEMS WITH COMPUTERIZED ADAPTIVE TESTING

This section discusses several classes of research problems inviting investigation in current and future CAT systems. It is organized to respond to some research problems and needs identified by Wainer and Kiely (1987).

Context Effects

Significant effects on item parameters and item performance have been shown to depend on the relationship with other items in the test (Eignor & Cook, 1983; Kingston & Dorans, 1984; Whitley & Dawis, 1976; Yen, 1980). In conventional testing, every examinee receives every item in the same order. Thus, the context effects are the same for all examinees. Because each examinee receives a tailored set of items in a tailored order in a CAT test, there is the possibility of differential context effects for different examinees. One potential solution to this problem is to conduct IRT calibration studies that counterbalance the pairings and sequences among all of the items in an item bank. The resulting parameter calibrations should average out the context effects. Research studies should also be conducted using repeated administrations that offer the same items in different sequences.

One kind of content effect is cross-information; that is, the correct or erroneous information that one item might provide concerning the answer to another item. To solve this problem, as in the conventional test-development situation, the item banks for computerized adaptive tests should be carefully checked by technical reviewers to remove any items that could provide cross-information to the examinee. Because it is virtually impossible to inspect all possible pairs in large item banks, technical review alone is unlikely to solve this problem, even at greatly increased review costs. Partial automation might help. New semantic search techniques by computer could be used to identify identical matches or synonym matches between any item stems or answers throughout the item bank.

Unbalanced Context

This problem arises when there is repeated emphasis on a particular content area or skill throughout a test, rather than balanced emphasis across content and skills. Adaptive tests make it difficult to maintain constant specification, which requires a balanced sampling of different content areas. One potential solution to this problem is to administer a computerized adaptive test with some additional domain specification criteria. For example, the College Board and ETS computerized placement tests (Abernathy, 1986) call for a test specification template within which the computerized adaptive testing is conducted.

Lack of Robustness

This problem occurs because the shorter computerized tests lack the redundancy of longer tests. The impact of an incorrectly functioning item is much greater in a short CAT test than in a longer, conventional paper-and-pencil or computerized test. One solution to this problem is to require a more stringent test-termination criterion in the computerized adaptive test (longer fixed test length, smaller standard error of estimate, or higher test-information values). Research has shown that computerized adaptive testing can reduce test length by 35%, perhaps by as much as 50% to 75% in some applications, while retaining the same precision. A partial solution, therefore, is to administer more items, reducing the advantage in test length in favor of increased robustness.

Item Difficulty Ordering

In conventional test development, a test is typically designed to sequence items from less difficult to more difficult. This allows almost all examinees to warm up with some success on easy items. Although this feature might increase fairness and validity, it also favors strategies for guessing on the basis of presumed difficulty, which reduces validity and is inequitable in favor of those coached in test-taking strategies.

One standard approach to CAT is to administer an initial item of average difficulty. This is optimal for neither high-ability examinees nor low-ability examinees. A potential solution to this problem is to provide short locator tests consisting of five to six items, which span the spectrum of item difficulties in the item banks. The locator test approach can insure a more accurate initial estimate than single items of average difficulty. It also offers some easy items for low- and high-ability examinees.

A second solution to this problem is to initiate the computerized adaptive test at a lower difficulty value (for example, at the 30 or 35 percentile value, rather than at the 50 percentile value). On the average, this would require slightly longer computerized adaptive tests, but it would insure that the majority of examinees would experience some moderately easy items. In all cases, CAT reduces the effects of coaching in test strategies based on the information that test makers place easy items first and difficult items last.

Wainer and Kiely (1987) suggest the investigation of "testlets," groups of items that carry their own context. Testlets might reduce the effects of context, cross-information, and sequence that occur in CAT.

Paragraph Comprehension Items

Wainer and Kiely (1987) note that most computerized adaptive testing systems have opted to develop and administer shorter paragraph comprehension items that present the text paragraph and multiple answers on a single screen. Research by Green (1988) shows that these short paragraph comprehension items are more similar to traditional word knowledge items than to previous paragraph comprehension tests, which changes the construct measured. A potential solution to this problem is to develop longer computerized text paragraph comprehension items. This will require higher resolution screens or the development of easy paging and prompting techniques that will help examinees know which page of a multiple-page item they are reading, how many more pages are in the text selection, and how to move quickly to a given text page.

Research should also be conducted to identify the impact of calibrating multiple questions associated with a standard paragraph comprehension item. Andrich (1985) has done some promising work in the calibration of multiple questions associated with a single passage. If such sets of questions are administered intact, there appears to be no bias in ability estimation.

Other Research Issues

Research on multidimensional IRT models has been initiated (Mislevy, 1987; Reckase, 1985), and should be continued. Research is also needed to further clarify the strengths and weaknesses of alternative ability estimation approaches (Bayesian, maximum likelihood, expected a posteriori, etc.). New test-construction procedures such as the testlet approach (Wainer & Kiely, 1987) should also be investigated.

Because examinees are tested at their level of ability, an adaptive test might be less boring for high-achieving examinees and less frustrating for low-achieving examinees. This claim has been made but not validated.

Additional research should be conducted with single and multiple test-termination criteria (fixed number of items, specified standard error, specified test information value, etc.).

THE THIRD GENERATION: CONTINUOUS MEASUREMENT (CM)

DEFINITION

The continuous measurement (CM) generation uses calibrated measures embedded in a curriculum to continuously and unobtrusively estimate dynamic changes in the student's proficiency. Tasks measured may be items, item clusters, exercises, unit tests, or independent work assignments. Changes may be observed in the amount learned, the proficiency on different tasks, changes in the trajectory through the domain, and the student's profile as a learner.

The differentiating characteristic of CM is the ability to specify dynamically a learner's position on the simple and complex scales that define a growth space. Continuous measurement produces a trajectory over time for the individual who is working to master a domain of knowledge and task proficiency. Measurement is accomplished by assessing the performance of each individual on tasks calibrated to serve as milestones of accomplishment. The milestones that make CM possible are embedded in a curriculum so that measurement is unobtrusive.

The CM generation assumes a two-part definition of curriculum: (a) a course of experiences laid out to help the learner grow toward certain educational ends, that is, a path through a domain; (b) a set of course markers, or standards, that serve as milestones of accomplishment along the way, that is, beginning, intermediate, and terminal markers.

The continuous measurement generation will not spring full blown from either the curriculum side or the measurement side of its parentage. Rather, it will grow slowly from its end points. Admissions testing, at the beginning, and certification, at the end of a course of study, have been well researched. As research and development progresses toward the center of a curriculum, rather than toward the end points, continuous measurement of progress can occur. Development of the CM generation will necessarily bring measurement scientists into the area of specifying or modeling more carefully the substantive content domains to be mastered. This will necessitate a major expansion of current practices in test specification (our current method of specifying domains) and augment these with methods of job, task, and cognitive analysis and knowledge-acquisition methods adapted from the field of expert systems. It will also bring measurement science more powerfully than ever before into issues of the construct validity of measures of cognitive and learning processes, and the measurement of change.

Associated with the definition of the CM generation are eight features, which generally follow the properties listed in the CM column of Table 1.

1. The computer requirements are those of a computer-aided education (CAE) system, with enough speed and capacity for CAT calculations. The CAE system is usually housed in a learning resource room where practice and assessment can take place. However, it is also possible to monitor responses in groups in regular classrooms using hand-held response devices.

2. Measurement occurs continuously, as it is embedded in the curriculum. Exercise modules themselves are calibrated in much the same

manner as items are calibrated in the second generation.

3. Continuous measurement is unobtrusive. Because measurement occurs when students' responses are monitored as part of their ordinary learning activities, testing does not stand out as a separate activity. Achievement testing occurs automatically as students work through the curriculum exercises. The testing of learning strategies could occur automatically as students choose different methods of approaching or avoiding the curriculum exercises and these choices are monitored. Other measurements of learner profile variables, including preferences and abilities, can be accomplished unobtrusively by inserting items into the curriculum at appropriate intervals. It should be noted that unobtrusive does not mean measurement occurs without informed consent.

4. Continuous measurement differs from the first two generations in emphasizing dynamic, rather than static, measurement primarily for individual purposes, not institutional purposes.

5. Data from continuous measurement should be available to, and of value to, learners and teachers alike. A representation of the domain of knowledge and expertise to be mastered is available for reference by the students and teachers. The progress of individual students on this representation, which might be called a mastery map, is also available for continuous reference.

6. The scaling in CM is more complex than that in CAT. Instead of the unidimensional scales of the CAT generation, CM deals with multiple and often multidimensional scales, but might summarize them into a single composite score or objective function, to track progress for each learner on a personalized mastery map.

7. Reference tasks are calibrated on interval scales of measurement in the CM generation. A reference task refers to, and might simulate, real-world or joblike performance requirements. Reference tasks generally are more complex than single items and require multiple responses. As learners encounter and experience these different tasks, a continuing estimate of changes in achievement (the learner's trajectory) can be estimated adaptively.

8. The CM generation will not be mature until research foundations have been established for the new psychometric procedures needed. This might involve extensions of IRT or new procedures entirely. Enhanced methods for construct specifications are needed to develop valid measures of learning outcomes at different stages of mastery. A useful set of learner profile measures is needed to characterize individual differences in learning while using such systems. Research is needed to establish implementation designs that will enable teachers and learners to become proficient over time in the new roles required for mixed forms of assessment and learning. Preservation of the conditions for validity and for the proper use of measurement depends on excellent implementation.

EXAMPLES OF PARTIAL CM SYSTEMS

No complete example of a CM generation system has been developed to date. However, several computerized educational systems have been developed that illustrate incomplete continuous measurement systems. These systems exemplify subsets of the eight properties that characterize CM. Two examples have been selected for discussion here. The first of them, the TICCIT system, was an ambitious CAE system developed in the early 1970s. It featured continuous and unobtrusive measurement, mastery maps, and a range of different reference tasks. TICCIT did not have calibrated reference tasks or scaling. The second example is a reading curriculum developed for the WICAT Computer-aided Education System. This reading curriculum illustrated most of the properties of a complete CM system, including calibrated exercises. A pilot study will be reported in which an attempt was made to scale the exercises. The CM concept has been developed further by researchers at ETS, under the designation Mastery Assessment Systems. The ETS concept is described herein.

The TICCIT System

TICCIT stands for Timeshared Interactive Computer-Controlled Information Television. The hardware for this system was designed by engineers at the Mitre Corporation in 1971 and 1972, and the instructional strategy, courseware, and instructional logic were designed by a group of instructional and computer scientists at the University of Texas and Brigham Young University. TICCIT was one of two major National Science Foundation funded CAE systems developed in the early 1970s. The other was the PLATO system, developed at the

University of Illinois.

The TICCIT hardware consisted of up to 100 color video monitors that could present mixed text and graphics in up to eight colors. Digitized audio was available, and videotapes could be switched to any terminal. Students responded with a typewriter keyboard when required to make a selection or to type a word or phrase. Most of their responses, however, were entered as single keystrokes on a special learner-control response pad at the right-hand side of the keyboard. This keypad had the keys MAP, OBJECTIVE, RULE, EXAMPLE, PRACTICE, EASY(er), HARD(er), HELP, and ADVICE.

TICCIT was initially designed with two courses in freshman and remedial mathematics and two courses in freshman English grammar and composition for community colleges. It was implemented at Phoenix College in Arizona and at Northern Virginia Community College in Alexandria. Educational Testing Service evaluated the project and found that TICCIT classes gained significantly more than control classes in both mathematics and English (Alderman, 1978).

Information about TICCIT can be found in Bunderson (1973) and Merrill, Schneider, and Fletcher (1980). An early history of CAE, including TICCIT and PLATO, can be found in Bunderson and Faust (1976). The TICCIT project assets were acquired by the Hazeltine Corporation after the NSF funding was completed, and the system was improved substantially during the subsequent 16 years. Its applications have primarily been in industrial and military training, rather than in education, in part because of the significant sociological and economic problems blocking implementation of computer-aided education in mainline college instruction. TICCIT has been acquired from Hazeltine by Ford Aerospace and it has continued as a viable CAE alternative.

TICCIT exemplified CM in that testing and measurement occurred unobtrusively whenever students were working at the terminal. Its psychometric theories were simple, and it did not have calibrated tasks or the particular research foundations discussed in point eight of the definition of CM. However, TICCIT provides a continuing example of the first five properties of CM. It exemplifies particularly well the mastery map concept, including a means of tracking student progress through a knowledge domain.

The TICCIT map was represented as a series of "learning hierarchies" (see Gagné, 1968). At the top level was the course map. In the English course, for example, the course map defined a set of units for learning to structure and organize writing and another set for learning to edit written compositions for grammar, mechanics, and spelling. Each unit on the course map was represented in further detail by a unit map, which consisted of a series of lessons, also arranged in a learning hierarchy. Each lesson also had a map consisting of a hierarchically organized set of segments. Each segment constituted a single objective, usually a concept, a principle or rule, or a kind of problem to be solved. The objective and different versions of a definition or rule statement were available for each segment. The segments included many practice items, classified as easy, medium, and hard, each with help available.

The instructional prescriptions for each kind of objective were developed according to what has since come to be called Component

Display Theory (Merrill, 1983; Riegeluth, Merrill, & Bunderson, 1978). Component display theory was a substantial elaboration of Gagné's earlier concepts of a prescriptive instructional approach, first promulgated in his book Conditions of Learning (Gagné, 1965). TICCIT, therefore, had a type of construct validity based on a taxonomy of learning objectives and it utilized the associated conditions for teaching each type of objective.

TICCIT maps provided a continuous display of progress, always available to the students on their individual video screens. The data on these maps were summarized for teachers in weekly printed reports. As a student progressed through a segment, the data for that segment box were updated, and the student was given feedback on performance. When all the segments in a lesson had been mastered, the student could take a lesson test. Each lesson test used a simple kind of adaptive testing logic called the Wald Sequential Testing Procedure, in which items were selected randomly or sequentially until a determination could be made as to whether the student had mastered or failed the lesson. Statistical calculations were made after each item. When the student was in an indeterminate state, further items were administered. Item administration stopped as soon as a determined state was reached.

TICCIT had the ability to track the progress of each student, even though students were moving at different rates through different sequences of lessons. Its lesson and unit maps with status information were a popular and successful feature. Menus, a currently popular interface convention, with the addition of continuously measured status information, would serve just as well.

Another continuous measurement feature of the TICCIT system was the advisor function. The complete advisor program as conceived was conceptually ahead of its time, because the artificial intelligence techniques needed to implement it properly were not available in 1972. The advisor was designed to employ measurements of the student's progress and use of different learning strategies and tactics to provide interpretive advice. The progress measurements would be used in connection with a set of prescriptive instructional rules to provide feedback to the learner when requested or when certain conditions were met, as calculated by the computer. Because of the complexity of programming such a system, by using decision tables instead of expert system and relational data base methods, the TICCIT advisor did not achieve its ambitious goals; indeed, it will be seen that such goals belong to the fourth generation of intelligent measurement applied in CM settings. Although the TICCIT advisor fell short of its goals, it did accomplish a major third-generation function. It provided each learner with a score indicating how well he or she was doing in the practice problems constituting a given segment. Students used this information frequently. There appeared to be a great thirst for dynamic information on how well an individual was doing.

The teachers' weekly reports of progress on the TICCIT lessons were also an important part of the implementation plan. Computer activity became more and more closely integrated with teachers' classroom activities as they learned to use the reports. Success with continuous measurement in educational programs like TICCIT depends heavily on an excellent implementation plan in which teachers learn new roles and students learn new habits and practices.

Calibrating Computer-administered Learning Exercises

Research conducted by the authors of this chapter at the WICAT Education Institute in 1980 provides an example of some of the benefits of calibrating exercises embedded in a curriculum, in this case the WICAT Elementary Reading Curriculum.

The WICAT Elementary Reading Curriculum. First developed under an Office of Education grant, this computer curriculum was designed as a practice and feedback system to accompany classroom instruction in reading comprehension. It consisted of a series of stories especially written for different grade levels, spanning the range of difficulty from the third through about the seventh grade. The system was based on a learner control philosophy, not unlike TICCIT.

Learners were allowed to select stories to read, within their own grade level, from a graphic map that resembled the front page of a newspaper or magazine. At a higher level of generality, and somewhat equivalent to the TICCIT course map, was a list of different newspaper or magazine titles. The equivalent to the TICCIT unit map was the "front page" with printed titles, frequently accompanied by a graphic identifying each story inside the "magazine." No status information was provided by these maps, except a record of stories completed. They functioned as tables of contents more than as mastery maps with status information. The students would select a story to read and then enter into an exercise involving that story. Each exercise consisted of reading a screen or two of information, answering interspersed multiple-choice questions, underlining key words on the screen to explain why a particular answer was given, and typing short answers. The typed answers were not judged by the computer. Self-corrective feedback was given, so that the learner could judge her or his own free response. These types of questions were repeated several times until the story was finished. Ten to 20 scorable responses were available for each story. A grading standard for each exercise had been established by the curriculum authors, analogous to the scoring standards within the TICCIT segments.

Manager programs were available to keep track of responses at a detailed level, resulting in a rich set of data for use by researchers, teachers, and students. Systems like the WICAT System 300, a 30 terminal on-line computer system for schools, or networked systems that tie microcomputers together, are needed for the centralized record keeping necessary in CM. The WICAT manager programs made research possible utilizing student response tapes from the children in the third through sixth grades at the Waterford School in Provo, Utah. As in the TICCIT system, the teachers had weekly reports showing student progress through the various stories at each grade level. The students had access to less information than the TICCIT advisor provided, but they were able to tell if they were doing well or poorly in each exercise. They had learner control that allowed them to respond to this information and move in and out of the exercises at will.

The WICAT reading curriculum also had an adaptive strategy for giving the students easier or more difficult reading exercises. When they passed a reading exercise at one grade level, the computer would move them to a harder grade level. The computer could also keep them at the same level or introduce them to easier exercises. The goal was to keep them at an optimally challenging level of difficulty. Thus, the WICAT reading curriculum involved continuous measurement, monitoring of progress, and some automatic adaptation of difficulty. Unfortunately, the teachers reported that the grade levels assigned by the computer in that early version of the courseware did not seem to work well. Students would advance into more difficult material too rapidly, creating problems with the quality of their work and with their motivation. The grade level parameters for each reading exercise had been established by the authoring team through the combination of Frye Reading Indexes and subjective judgments.

Cross-fertilization between instructional and measurement research occurred naturally at this time because the work in adaptive testing and studies of the reading data occurred simultaneously. Two of the authors were also involved in the Armed Services Vocational Aptitude Battery Computerized Adaptive Testing that WICAT Systems was conducting for the Department of Defense (Olsen, Bunderson, & Gilstrap, 1982). They had implemented CAT tests for ASVAB on a specially designed hardware system and were using the BILOG program (Mislevy & Bock, 1982) to calibrate items. It was thought that the curriculum exercises themselves could be calibrated as easily as single items and that the difficulty parameters obtained from the calibration would provide an excellent scaling for the reading exercises, potentially superior to the curriculum authors' judgments of reading difficulty and associated grade levels.

These considerations led to the design of a pilot study wherein the ratios of correct to attempted (using the first attempt in each reading exercise) would be used as a response vector to calibrate the exercises. This pilot study was seen as the demonstration of a concept: the potential benefit of equal-interval scaling of curriculum exercises. It was not seen as a finished methodology, because adherence to the assumptions of IRT could not be assured, nor were the samples as large as would be desired for each exercise. There was, nevertheless, reason to believe that the reading stories were sufficiently independent of one another so that the assumption of local independence was not unreasonable. The assumption of unidimensionality was not unreasonable for reading items, but it was not checked. Because of learner choice, the stories were taken in a quasi-random order. The correct-attempted ratio of each student's performance within each reading exercise was more reliable than single-item responses.

The response vectors from the reading curriculum were calibrated using the BILOG program. As expected, the difficulty (theta) parameters showed that the grade level designations of each story in the curriculum were not supported by the empirical data on story difficulty.

Some grade level jumps rated large by the reading curriculum authors were in reality tiny steps in difficulty, whereas exercises presumably at the same grade level provided giant steps in empirically determined difficulty. The use of the calibrated difficulty parameter offered the curriculum developers an opportunity to make a substantial improvement in sequencing the on-line reading curriculum.

Table 3 compares the empirical difficulty parameters from the BILOG program for 48 argumentation stories, with the judged readability indexes for each story (WICAT, 1983). The correlation between these two difficulty estimates is very low (r = -.07). Clearly, tasks requiring learners to think and respond are different from the difficulty (readability) of a reading passage itself. Note that the two extreme stories (5 and 14) on the IRT scale both had a judged grade level of 7. Story 14 had a difficulty value of -.9, very easy, whereas story 5 had a difficulty value of 2.0. Notice, also, that the 6th and 45th ranked stories on the IRT scale both had a judged difficulty of 4, the easiest grade level for the argumentation exercises. (Only the argumentation exercises were calibrated. Other reading exercises were less difficult, and teachers reported fewer problems with the computer adaptive strategy.)

The scaling of exercise units has considerable promise, once the proper psychometric procedures can be developed for a wide class of exercise types (including hierarchial and cumulative curricula). It can provide valuable feedback to the curriculum developer. This feedback could be very helpful, for example, in examining the details of exercises that prove to be much harder or easier than expected and in developing an effective adaptive strategy for moving the students along. It can also provide a continuous measurement of achievement in reading comprehension for monitoring progress by using a mastery map. The standard scale values of the BILOG difficulty parameters can be converted to a grade-equivalent scale or normal curve

TABLE 3

COMPARISON OF IRT DIFFICULTY PARAMETERS WITH JUDGED GRADE LEVEL IN A SET OF COMPUTERIZED READING EXERCISES

STORY NO.	IRT DIFFICULTY PARAMETER	IRT RANK ORDER	JUDGED GRADE LEVEL (READABILITY)
5	1.98	1	7
34	1.86	2	5
35	1.82	3	5
12	1.79	4	7
32	1.65	5	5
26	1.31	6	4
1	1.28	7	7
29	1.23	8	4
27	1.21	9	4
3	1.11	10	7
7	1.11	11	7
2	1.10	12	7
28	1.08	13	4
44	1.02	14	6
13	1.01	15	7
4	.96	16	7
18	.92	17	6
16	.85	18	6
41	.81	19	4
43	.80	20	6
48	.77	21	4
42	.72	22	6
21	.71	23	4
39	.66	24	6
6	.61	25	7
31	.60	26	5
38	.60	27	6
45	.58	28	6
36	.58	29	6
47	.57	30	4
24	.55	31	4
33	.55	32	5
20	.53	33	6
37	.52	24	6
22	.40	35	4
19	.39	36	6
10	.36	37	7
23	.35	38	4
25	.34	39	4
46	.31	40	6
15	.26	41	7
17	.22	42	6
9	.20	43	7
11	.14	44	7
30	-.04	45	4
40	-.56	46	6
8	-.57	47	7
14	-.93	48	7

Rank Correlation = -.07

Note: It is possible for a calibrated set of curriculum-embedded exercises to serve as a standardized test of reading comprehension.

equivalent scale. When there are multiple dimensions, a cumulative summary can be obtained by developing a function that summarizes progress toward the multiple objectives of the course of study.

The pilot research with the WICAT reading curriculum provided a more complete example of what a continuous measurement system might be than had the TICCIT investigations. This example shows that measurement science has much to offer to education, both for curriculum development and for delivering curriculum intelligently and adaptively. Continuous measurement holds the promise of providing unobtrusive, frequent, reliable, and valid data. The continuous and sequential nature of the tasks anchoring the measurement provides many opportunities for continually assessing and improving reliability and construct validity.

Preliminary Evidence of Construct Validity

Some evidence of construct validity was obtained by relating the IRT scores to standardized test scores. IRT estimates of students' ability were obtained from the individual response vectors of each Waterford student, coupled with the parameters of the exercises the students passed and failed. Reading comprehension scores on all of the students at the Waterford School were also available from the Iowa Test of Basic Skills. The individual ability levels estimated from the calibrated curriculum exercises were included in a factor analysis with scores from the Iowa Test reading comprehension subscores. The factor analysis showed that the continuous measurement scores loaded significantly on the factor represented by the four reading subscales of the Iowa Test of Basic Skills. The factor loadings of the Iowa subscales ranged between .79 and .86, whereas the continuous ability estimates from the computerized curriculum had significant, but smaller, loadings ranging between .24 and .45 on the same factors. The authors took these results as evidence that similar constructs were measured by the paper-and-pencil reading exercises and the computer- administered reading exercises, but that method-of-measurement variance was also present.

It is unlikely that factor analysis will prove the most useful tool for determining the construct validity of measures obtained through continuous measurement. Construct validity can and must be approached using a variety of correlational and experimental methods. Many new options exist for mixed correlational and experimental approaches to understanding the constructs that constitute a domain of knowledge, when the associated curriculum has an embedded continuous measurement system. Measurement science alone is not enough, however. An interdisciplinary synthesis is needed between the cognitive and instructional sciences. The emerging scientific foundations for construct validity and dynamic testing are discussed later in this chapter.

Differences in Utility Between CM and School Achievement Tests

Turning from the topic of construct validity to the topic of utility represents a strong move from basic to applied questions, from theory to pragmatics. The CM generation is intimately involved in pragmatics, as well as theory, because education and curricula are pragmatic subjects. Utility of measures deals with their practicality and ease of use in live educational settings. By becoming unobtrusive, CM takes a giant leap in utility.

The procedures for obtaining estimates of individual student proficiency on calibrated curriculum-embedded tasks are very different from those of a nationally standardized paper-and-pencil test. The contrasts between these procedures are useful for highlighting the benefits in utility of the third generation for learners and teachers.

In the case of the standardized test, "testing days" are an obtrusive intervention into the school week. For example, at the experimental Waterford School, where the pilot reading study took place, the headmistress would announce that testing would take place during a certain week in the spring, and the teachers would solemnly pass this information on to the students. Such testing had not been enjoyable in the past, and the news would be greeted with moans and groans from the students. Testing is a traumatic experience for most of them. On the testing day, test administrators entered the classroom, instead of the familiar teachers, and strove to create a highly standardized environment. They introduced careful timing for subtests, controls against cheating, and instructions on how to fill out the answer sheets or test booklets. The students knew that there would be a formal report and that it was very important to do well. Students might have been ill or emotionally upset that day, but they had no choice about the test date.

Administration of the computer curriculum with unobtrusive continuous measurement is strikingly different from group paper-and-pencil testing. At the beginning of the year, the teachers introduce the students to the computer room and establish ground rules. During the year, students go to the computer room happily, sometimes insisting that the teacher break off classroom activities so they won't miss any

BEST COPY AVAILABLE

time. In the computer room, they settle down quickly to work at their own paces and at their own positions within the reading curriculum. Measurement takes place day after day, and the cumulative accuracy of the estimates of reading ability and its rate of change increases to higher and higher confidence levels. Given unidimensionality and an appropriate scale, the students' trajectories from easier to more difficult reading exercises can be tracked, plotted, and reported to teachers and researchers at frequent intervals.

MASTERY ASSESSMENT SYSTEMS AS CONTINUOUS MEASUREMENT

The concept of mastery assessment systems, a CM generation concept, was developed during 1986-1987, by ETS researchers (Forehand and Bunderson 1987a, 1987b). By mid-1987, several multiyear research projects had been initiated to develop mastery assessment systems.

Two features of mastery assessment systems can be noted at the outset. First, a mastery system has a role to play in curriculum planning, but is not itself a curriculum. Second, the term mastery does not refer to minimum competence alone.

The developers of a mastery assessment system should go to some effort to map elements of the defined domain of knowledge and expertise into the goals of a variety of localized curricula. They should identify generally accepted milestones of learning in a particular domain, covering a particular level of learning and extending below and above it. They should calibrate these measures to make sure that they provide a smooth and continuous series and do not embody large jumps in difficulty or complexity that would trap many learners at a certain level. It is the task of local educational jurisdictions to develop or select curricula. As a part of this responsibility, they could select a subset of measurable milestones from the larger sets of the MS and embed them in their own curricula as a measurement framework.

When a measurement organization obtains group consensus on learning milestones in a particular subject area, that consensus usually converges on what might be called a minimum competence standard. The mastery that is assessed in a mastery system, on the other hand, looks forward to a time when, after long commitment and effort, the learner has obtained a lifelong capability.

Mastery signifies achievement of personal learning goals that go beyond minimum competence. Mastery is personal and unique and is achieved after long periods of persistence and commitment. The assessment of higher levels of mastery must involve unique productions (e.g., complex problem solving, oral presentations, written analyses, portfolios). Some of the precursors of mastery can be assessed at earlier levels of growth, by encouraging the learner to practice some element of mastery appropriate to his or her growth stage. At intermediate stages of learning, the student can thus experience what it means to persist and to expand upon what has been learned until able to do something and know something really well.

Assessment includes the use of standardized measures of competence and guidance for judging the precursors of mastery at various levels. This guidance can include disciplined subjective scoring or, in the future, intelligent computerized scoring. Instructionally sensitive assessment will have new properties and paradigms not fully developed by a measurement science built to support certification, selection, and classification.

Components of a Mastery System

A mastery system would require a CAE system, as described. It need not be as elaborate initially as TICCIT or the WICAT system 300. Though the assessment systems could partly be implemented on paper, at least one computer for scoring and record keeping would be necessary. Major nonhardware components of a mastery system include the following:

1. Mastery map usable by learners and teachers to envision and communicate about learning goals
2. Reference tasks
3. Calibration of items and reference tasks
4. Instruction-oriented scoring system for each reference task
5. Professional development program for teachers

The first four of these components depend on measurement concepts; the fifth is an implementation concept. These components are intended to serve instruction; therefore, they could be linked to instructional components. Instructional components would include repeated practice in reference tasks, subscoring to guide the instructor in coaching, and report-generating systems for students and teachers. The term coaching is meant to be an analog to the instructional process that an excellent athletic coach uses. This might include modeling the desired performance, observing practice trials, prompting, encouraging, and fading the prompts as the performance becomes adequate.

The Mastery Map. In the development of a mastery system, responsible educational officials would work with measurement experts who are competent in CM to embody a selected subset of calibrated reference tasks as the markers, or milestones, needed within their own curricula. This plan would be visualized as a mastery map that would give the learners and teachers an overview of "the journey at a glance" at the beginning of learning. The mastery map would also permit communication about initial placement and about next steps in accumulating progress. The mastery map could be visualized on a large wall display for all, but individual maps with status information should be made available graphically on the computer, as exemplified by the TICCIT map displays.

Reference tasks. A reference task is generally more complex than a single item. It might be a testlet, as defined earlier; a curriculum-embedded exercise requiring multiple responses; or a simulation exercise. A reference task is contextualized. It refers to some real-world work that communicates to students, parents, and community the relevance of the things being practiced. A reference task might also refer to component process constructs important to the mastery of the task and useful in coaching. A record of an individual's accomplishment on reference tasks can build up the self-confidence of the learner. Table 4 contrasts test items and reference tasks.

Table 4
Test Items Versus Reference Tasks

	Test Items		Reference Tasks
1.	Usual administration is by paper and pencil.	1.	Usual administration is by interactive computer.
2.	Written objectives prescribe test items.	2.	Flowcharts and interaction specifications prescribe reference tasks.
3.	Each item requires a single response, usually multiple choice.	3.	Each task requires multiple responses, which together provide for a quantitative assessment of degree of success.
4.	Scoring is dichotomous a. Pass b. Fail	4.	Scoring is trichotomous. a. Pass (competence demonstrated) b. Needs coaching and practice c. Not ready for this task
5.	A complete test with subtests can be used for diagnostic purposes.	5.	Simultaneous subscores are taken to measure component processes and states, which provide data to guide coaching.
6.	Items and entire tests are often decontextualized. Learners, parents, and community figures might not see the relevance of the question to valued capabilities in the real world.	6.	Tasks refer to or simulate aspects of valued real-world activity (e.g., in college or a job).
7.	Items can be calibrated and placed on a measurement scale.	7.	Reference tasks can either be calibrated into the same scale as items or positioned on a contrived growth-objective function, with different regions representing stages of mastery.
8.	Except in CAT systems, administration of next item is fixed by its order on the page.	8.	Next response request is determined dynamically.
9.	Practice uses up test items after one attempt, making them of little value for repeated practice.	9.	Some reference tasks require files of alternative stimuli for practice (e.g., paragraphs to read), but many of them, including a simulation or gamelike event, can be practiced repeatedly without using up material.
10.	The objective and specification of how an examinee would succeed are neither suitable for learners to view, nor are they now presented.	10.	A model of mastery can be made available to help learners see how it should be done. Contextual referencing makes modeling more realistic.

Calibration. Reference tasks can be placed on scales to show the degree of growth they represent. Test items, perhaps grouped into clusters, or testlets, can also be placed on such scales. For example, the following tasks were used in the NAEP study of the literacy of young adults (Kirsch & Jungeblut, 1986). They assess literacy skills used in interpreting documents. The scale values are statistically determined measures of difficulty based on IRT. In the literacy study, they are also described and explained in terms of task features that account for variations in difficulty.

Sign your name on the line that reads "signature."

Scale value: 110

Put an x on a map where two particular streets intersect.

Scale value: 249

Fill in a check to pay a particular credit card bill.

Scale value: 259

Use a bus schedule to answer: On Saturday morning, what time does the second bus arrive at the downtown terminal?

Scale value: 334

Use a bus schedule to answer: On Saturday afternoon, if you miss the 2:35 bus leaving Hancock and Buena Ventura going to Flintridge and Academy. how long will you have to wait for the next bus?

Scale value: 365

These examples illustrate the calibration of reference tasks in a mastery system. Calibrated values could reflect both the educator's or the expert's analysis and the empirical results. Once the reference tasks in a set are calibrated, the scale values are given meaning by demonstration of the constructs of knowledge and skills required to succeed at tasks with a given range of values. Each mastery system would have its own calibration. The scale and scale interpretation would be developed and validated for a particular content, level, and purpose.

In most cases, educational growth is not linear and additive. When experts are compared to novices on a wide variety of tasks, they are characterized not so much by quantitative increases in the amount of knowledge, as by differences in the perceptual and conceptual organization of knowledge. Experts, as contrasted with novices, organize knowledge hierarchically. They group information according to underlying principles, have easier access to the information they have stored, and use information more flexibly. Therefore, *calibration is a matter, not of adding up units of learning, but rather of determining indicators that mark progression along a continuum of growth from novice to expert*. Sharp increases in difficulty in a calibrated sequence are often a signal to perform a deeper cognitive analysis.

Educators use such terms as *novice*, *advanced beginner*, *competent*, *adept*, and *expert* to describe variations in growth. Calibration of reference tasks gives such descriptors meaning in terms of statistically determined scales and conceptual analysis of the properties of tasks at each scale position. It is expected that indicators and models of growth differ for psychomotor skills and for academic knowledge; for children's initial acquisition of academic skills and for professionals' acquisition of new knowledge; for learning science and for learning a foreign language. One of the challenges for research is to identify underlying principles that will permit a comparison of mastery systems across domains and to provide new guidance, and automated systems, to aid in developing such mastery systems. A challenge for developers of a new mastery system is to develop and justify useful and construct-valid calibrations for a given application.

Instruction-related scoring of reference tasks. Items are normally scored dichotomously, as correct and incorrect. It is possible to score reference tasks more finely, to connect performance with instructional strategy. Students might be placed in one of three categories: those whose performance demonstrates competence achieved, those who are in need of and ready for practice, and those who are not ready for practice on a given reference task. The development of scoring algorithms to make these classifications is based on a combination of expert judgment and systematic observation of the performance of students known to be proficient, as compared with those at a lower level. The scoring algorithms are based on the occurrence of correct responses and the nonoccurrence of particular responses that indicate misconceptions.

A professional development program for teachers. A mastery system is always accompanied by a professional development program for teachers, because it makes new professional roles possible in several ways. It frees some time for professional activities while learners are practicing on reference tasks. It provides a tracking system, so that teachers can make professional decisions about how to manage the progress of a class and its individual members. More advanced systems could provide information to aid in coaching individuals and groups. Teachers learn new practices in relation to these particular technological tools involving a mastery system, so that they can be successful at using the system for placement, tracking, classroom management, and, ultimately, for individual and small-group diagnosis and coaching.

A mastery system is designed to function within a community of learners and teachers, and users must learn to build and sustain such environments. A mastery system provides the opportunity to build and support a cooperative community of learners whose goal is to help and encourage one another, to teach one another, and to facilitate the maximum amount of learning for students and teachers. Professional development would include training in appropriate use and interpretation of measures and in methods to build and maintain appropriate climates.

How Mastery Systems Can Serve Individual Learners and Teachers

Mastery systems offer extended ways in which measurement can serve individual learners and teachers. Traditional test use has often emphasized such institutional purposes as admission, certification, job placement, grading, and classification. There have also been consistent efforts to serve individuals, such as through

Counseling and guidance
Advanced placement
Special recognition
Placement in a learning program
Diagnosis of learning problems
Self-assessment for personal knowledge and growth

Mastery systems multiply these opportunities by focusing measurement on growth in skill and knowledge. The goal of a mastery system is the advancement of students toward mastery. It provides data to help and ways to use the data. Table 5 lists 12 services that mastery systems offer to individual learners and their teachers. Early mastery systems are likely to address only a subset of these goals. Systems that succeed in meeting substantial subsets of these goals could provide successive new generations of services to learners and teachers.

Table 5

Mastery Systems: Possible Services to Individual Learners and Their Teachers

1. Initial Placement on the mastery map

2. Tracking within a well-defined map of competence and mastery to show current position, nearby options, achievement to date, and potential growth

3. Repeated practice on interesting and informative reference tasks

4. Trichotomous scoring systems for reference tasks to classify learners' current attempts as

    a. Fully demonstrating competence or mastery
    b. In need of coaching and more practice
    c. Not ready for the task

5. Presentation to learners of informative models of mastery (how successful students think and perform) relevant to particular reference tasks and mastery levels

6. Simultaneous measurement of component processes and states during reference task practice

7. Data to guide coaching based on component processes and states

8. Data to guide coaching based on metacognitive heuristics and strategies

9. Presentation to learners of information about their characteristic learning profiles

10. Data to guide coaching based on learner profiles interacting with coaching needs

11. Prediction of learning decay to prescribe review

12. Analysis of group records to facilitate classroom management

    a. To adapt rate of progress and depth of instruction for the whole group
    b. To select subsets of learners for small-group coaching
    c. To identify individuals in need of personal attention

THE ROLE OF LEARNER PROFILES IN THE CONTINUOUS MEASUREMENT GENERATION

The use of a profile of scores descriptive of different learner styles, strengths, and weaknesses has long been a dream of educators and behavioral scientists. The WICAT learner profile is an attempt to achieve that goal. Initiated by the authors at the WICAT Education Institute in 1983 and 1984, it consisted of a battery of computer-administered tests (CT and CAT generations) designed to profile the styles, abilities, and preferences of individual learners. The learner profile battery has been a source of illustrations for CT and CAT tests in this chapter, and batteries like it could become a major tool in educational measurement. Learner profile batteries in a CAT system place a rich set of learner profile scores in the context of a substantial body of computerized curricula.

As with earlier attempts to introduce highly individualized levels of measurement into educational and training settings, there are many pragmatic obstacles to widespread use. Funding for the necessary research is always hard to obtain. Potential users are frequently resistant to change. Thus, the progress in practical and scientific matters possible with such potentially powerful new systems is elusive.

The introduction of a substantial battery of learner profile scores into a school brings with it a major problem in data interpretation for teachers and students. Instruction in how to use the scores to improve learning is essential if users are to achieve the goals of their schools. Seeing one's profile as a learner is very informative and interesting to individuals, and it could well help learners become more confident, self-accepting, and effective. It might also help teachers be more accepting and sensitive to differences, if they are taught how to avoid abuses. Unfortunately, a curriculum dealing with constructs about individual differences currently has no place in the school's schedule. Only if it can be shown to aid in achieving conventional academic goals can a learner profile presently be justified.

Effective use of computer-aided education and mastery assessment systems, and the proper use of a battery of learner profile scores will require a substantial and lengthy professional development program for users. Curricula for such a professional development program cannot yet be defined in a way likely to achieve wide acceptance because there is no agreement among experts as to which variables are most important in a learner profile, let alone how to interpret and use prescriptively the variables that are better known. The field of education might have to wait until experience with learner profiles accumulates at user sites involved in CAE and CM before an expert knowledge base about effective use of such information can develop. Perhaps a prerequisite for development of such expert knowledge is the existence of CM systems described earlier. Until it is possible to measure individual learner growth trajectories, it will be difficult to evaluate alternate uses of learner profile data in improving the progress of learning in real educational or training settings.

New Learner Profile Variables

A new class of learner profile measures may emerge from the continuous measurement generation. These new variables might be more readily understandable and usable by educators. An example is taken from the WICAT reading curriculum. In a study of the response protocols from that curriculum, it soon became apparent that there was a wide range of different strategies for approaching the learning exercises. Students were given much choice as to which reading selections to study and the option to jump in or out of the exercises before completing them. They had knowledge of the pass-fail scoring on the exercises and could judge how well they were doing. Two extremes in learner strategies were observed. At one extreme were those students who were not troubled by initial failure. They were aware that they could attempt the exercise over again and that initial failure would not become a part of their permanent record. These students quickly tried the exercise, often failed it, but learned what was required in the process. Then they would go back through the exercise more carefully and pass it. This was more than memorization, because specific correct-answer feedback was not given. At the other extreme were students who assumed a record existed and refused to allow any type of failure which might blemish it. These students would escape from the exercise and sometimes not reenter it if there were any indication that they might not pass it on the next attempt.

Various intermediate strategies were defined by the researchers, including fail-pass, escape-pass, fail and avoid, escape and avoid. Discussions with the teachers brought out the fact that some of these patterns seemed characteristic of the students' performance in classroom activities. At the most serious extreme (the escape and avoid extreme) were some students who were quite timid and unsure of themselves. Other students at this extreme, however, were merely trying to avoid work.

Such students appeared attentive while at the computer terminal (or in the classroom), but they were not actively engaged in the learning process. The ability of the computer system to define and quantify this new class of strategies, observable because approach and avoidance can be measured, can have important implications for helping a variety of students.

Continuous Measurement of Learning Preferences

With systems suitable for continuous measurement, the curriculum can have a variety of options, such as visual versus verbal, structured versus holistic, or sequential versus simultaneous approaches to lessons. By choosing an option, students reveal their learning preferences. An ambitious attempt to provide learner control of different presentational formats was a feature of the TICCIT project. Students were given options to look at more visual, versus more verbal and teacher-like, explanations of the definitions or rules for each concept or principle. They could examine a selected range of examples, work more or fewer practice exercises, and look at different versions of "Helps" as they worked the practice problems or examples.

The TICCIT concept of learner control of strategies and tactics was a good one, but it needed expansion into additional types of learning components. Current and future CAE systems, with color graphics and audio options, could make additional kinds of learning components possible. Continuous measurement based on voluntary choices among these components would provide teachers with knowledge of the preference profile of each individual, and also of the group as a whole. Teachers could then present different parts of their own lessons in different ways, appealing to different profiles in the process.

In addition to the measurement of learning preferences through voluntary choices of on-line options, standardized and calibrated preference questions can be introduced at strategic points in the curriculum. The use of preference data can become a viable and useful part of the teaching and learning process.

Emerging Scientific Foundations for the Third Generation

In a real sense, construct validity is the fundamental scientific position for all measurement. The challenges of dynamic measurement demand that new solutions be found to the problems of construct validity of learning measures. Messick's chapter in this volume draws the inescapable conclusion that all validity concepts boil down to construct validity. There is, unfortunately, no simple and unitary set of procedures that can assure construct validity. The challenge of construct validity is to infer invisible constructs of human expertise from observable behaviors. The addition of certainty to this process is accomplished by testing as many inferences as possible, when the inferences are drawn from an understanding of the invisible constructs and their relationships to external behaviors. This might take researchers into such diverse realms as perception, learning, problem solving, and personality, and it will require cross-disciplinary cooperation.

For the third generation, the disciplines that appear most relevant at this time are the cognitive and instructional sciences. Embretson (1983) discussed the need to use models derived from cognitive science to represent the constructs involved in tests. She discussed how we need to link the construct representations to measures of individual difference that correlate with other measures of interest and value in applications. The ability constructs that have served as guides to psychometricians for many years are shown to be decomposable into functional components which may be described as representations of cognitive constructs. A test item can function in different ways and can effectively measure different cognitive constructs, depending on the test takers' positions on the component cognitive constructs that comprise an ability.

Embretson (1985a) applies this view of the construct validity of tests to the problem of test design and shows how component latent-trait models can be used in the test-design process. She shows that test designers can use three levels of cognitive variables (stimulus features, components, and strategies) to predict the difficulty of items and to determine the meaning of a test score in terms of the cognitive components and strategies the test items call upon. To gain this control over stimulus features, components, and strategies, test designers need subtask data that will enable them to determine which strategies or components are significant in different tasks. Subtask data is also necessary to diagnose individual test takers in terms of their use of components or strategies. As Embretson points out,

> Computerized adaptive testing can estimate ability by administering fewer items to each person, thus giving time for other tasks. Furthermore, the interactive format of computerized testing makes subtasks quite feasible, thus component latent trait models may have wider applicability

in future testing. (p. 217)

The analysis of cognitive components and strategies underlying tests of cognitive ability is now well advanced. Shephard and Metzler (1971) provide an important landmark in identifying the cognitive processes involved in the mental rotation of three-dimensional objects. Sternberg and his colleagues are building a systematic basis for componential analysis in measurement (see, e.g., Sternberg, 1977; Sternberg & MacNamara, 1985).

It is a different matter, however, to bring psychometrics and cognitive science together in the third generation than in the second, because dynamic measurement is required. Embretson (in press) explains that significantly different new developments are needed in the field of psychometrics to accommodate the realities of correct cognitive processes, erroneous cognitive processes, and the dynamically changing nature of these processes during learning. In this article, Embretson also reviews psychometric considerations for dynamic testing and shows that some views commonly held by psychometricians stand in the way of progress. Progress will require substantially different models and ways of thinking than have been sufficient for the static measurement of ability constructs.

Other researchers have dealt with the issues forced upon the field of educational measurement by progress in cognitive and instructional science. Glaser (1986) summarizes several challenges that cognitive and learning theories have raised for psychometricians. He outlines objectives that measurement models should consider if they aspire to deal with the domains of learning and instruction. Tatsuoka (in press) has developed a promising psychometric model that integrates item response theory with cognitive diagnosis. Her contribution is based on a clear understanding of the implications of cognitive science for construct validity and the need for latent-trait models that will avoid certain philosophical and scientific problems inherent in the application of current models. Her Rule Space model is applicable to the diagnosis of the status of learners on a set of correct and incorrect constructs of cognitive processing. This model treats the latent trait as a quantitative variable, not a categorical one. A test developed with Tatsuoka's procedures would yield diagnostic information about the probability of certain errors. Such information could lead to instructionally useful prescriptions.

Intelligent tutoring systems (Sleeman & Brown, 1982) offer a considerable challenge to psychometricians. Such systems provide models of underlying constructs that constitute expertise in a variety of subjects. Ideal, or expert, models are frequently accompanied by "buggy models" of incorrect procedures used by novices or students in the process of becoming more expert. These researchers are more interested in what is going on in the minds of learners than in how much of a quantity that might be scaled is being demonstrated. Therefore, there is a large gap to bridge between this work in artificial intelligence and cognitive science and the work of measurement scientists interested in dynamic measurement. Two recent books written to encourage dialogue demonstrate clearly that the gap is far from closed (Freedle, in preparation; Frederiksen, Glaser, Lesgold, & Shafto, in preparation).

RESEARCH ISSUES IN CONTINUOUS MEASUREMENT

Research issues in continuous measurement are too numerous to discuss in detail. The move from static to dynamic measurement and from a controlled testing setting to a continuing and complex educational program produce research questions at many levels of measurement science, cognitive and instructional science, behavioral science, and computer and information science.

There are many psychometric issues. One is the question of how to define the fungible unit of measurement. If it is no longer a test item, how do we define a testlet or a reference task? How do we scale and calibrate such entities? What about the assumptions of unidimensionality, local independence, and fixed, instead of moving and changing, proficiency? Another psychometric issue involves scaling and use of latency information.

Again, issues dealing with the cognitive and instructional sciences abound, and the continuous measurement environment provides a new instrument of vision for revealing to researchers the set of processes involved, within and across individual students, in performance on particular items or reference tasks. As learners progress from one level of proficiency to another, the evolution of these processes becomes visible through continuous measurement. Another question deals with where, in a mastery map, are the ranges of proficiency extending from novice to expert? Sharp increases in the difficulty of calibrated tasks might signal places to look for interesting changes in underlying cognitive structures.

Continuous measurement introduces issues of human development, group organization, group management, interpersonal relationships, and individual differences in group functioning. It provides new dependent and independent variables to enrich these studies.

The processes of change in the introduction of new forms of education using new tools and the products of science and technology constitute an important field of research. Without research-based design principles for introducing the change in bite-sized chunks, and in-service training over long enough periods of time, continuous measurement systems will not achieve their promise. Implementation research is likely to involve a variety of social science disciplines including anthropology, sociology, economics, and organizational behavior.

The computer and information sciences are obviously fundamental to progress in the field of continuous measurement and computerized instruction. Advances in hardware and software can have a profound effect on the cost and capabilities of the subsystems involved in CAT systems, continuous measurement systems, and computerized education alternatives. In the next section we address another powerful contribution of the computer and information sciences: the impact of fifth-generation computing on educational measurement. The field of computer science has used a generational framework for many years that should not be confused with the generational definitions for educational measurement presented here. In the remainder of this chapter, we deal with knowledge-based expert systems and a touch of natural language processing, and do not deal with some other advances (e.g. computer vision, robotics, speech recognition) also attributed to the fifth generation of computing.

THE FOURTH GENERATION: INTELLIGENT MEASUREMENT (IM)

DEFINITION

Intelligent measurement is defined as the application of knowledge-based computing to any of the subprocesses of educational measurement. The term knowledge based computing is used here, rather than the more familiar term, artificial intelligence, to draw attention to the notion that the knowledge and expertise of measurement professionals can be captured in a computer memory in a symbolic form called a knowledge base. This knowledge can then be used to replicate, at multiple sites through a computer, the expertise of humans who are otherwise restricted to one site at any time. Thus, with the aid of the intelligent computing system, less expert humans can perform measurement processes that require considerably more knowledge and experience than they presently have. Educational measurement is a knowledge-intensive discipline, and the knowledge is not commonly found among practitioners in education. Intelligent measurement introduces the ability to package knowledge, to replicate it in the form of a computer system that can interact with the user as an expert consultant or advisor, and to disseminate the expertise to many sites. It offers the field of measurement a powerful new way to bring the benefits of measurement to many educational practitioners who otherwise would have no opportunity to apply advanced methods in a knowledgeable fashion.

The fourth, or IM, generation can be contrasted with the others in terms of the properties summarized in Table 1. It assumes the existence

BEST COPY AVAILABLE

of a computer equipped with knowledge-based computing features, in either hardware or software. It also assumes that, through accumulated research and experience, it has been possible to capture symbolically expert knowledge and incorporate it into a computer as a knowledge base. For example, one type of knowledge base makes possible intelligent interpretations or prescriptive advice. Another type makes possible the automatic replication of complex scoring requiring human judgment. The first type of knowledge base models the expertise of counselors for applications in any generation involving interpretive comments about an individual's scores from a battery of measures. For CM generation applications, it models the knowledge of excellent teachers who are familiar with the subject, with the instructional system, with sound pedagogical practices associated with the system, and with knowledge of how to relate instructional alternatives to different learner profiles and trajectories. Another type of knowledge base (automatic holistic scoring) represents in the computer's memory the consensual knowledge of standards for mastery of certain reference tasks and the consensual scoring knowledge of experts in a subject domain.

As stressed in the discussion of CM, no computer-generated advice with important consequences for the individual should be used without scrutiny by an appropriate professional. The advice should come as a set of two or three alternative interpretations or prescriptions so that the user, guided by a professional for critical issues, could select or modify one.

A major difference between the fourth and earlier generations in the automation of test-administration processes is in the capability for sophisticated interpretation of measures, both static measures and measures taken during a dynamic educational process. This capability is only partially available through the third generation computer system and available only through expert people in the earlier generations. Intelligent interpretations of a given profile of scores are now available in many application areas; however, validated expert knowledge does not yet exist for prescriptive advice in CM generation applications.

The definition of IM given at the beginning of this section is general. This packaged intelligence can be added to computer programs designed to augment the work of users involved in any of the processes of educational measurement. A computer application program would perform a function like developing certain complex items, scoring them, or analyzing them. In each case, the application program could be accompanied by a knowledge base of expert decision rules and a data base of facts. This symbolically represented knowledge would be used by an "expert consultant" or advisor to guide the user in making informed decisions in the process of using the application program. Examples include:

Test Development Processes

Computer tools for job and task analysis, with advisor

Computer tools for developing test specifications, with advisor

Item and test development programs, with advisor

Test Administration Processes

Intelligent administration of individually-administered tests, with advisor to guide the paraprofessional

Natural-language-understanding expertise for scoring constructed responses

Interpretation of profiles

Intelligent tutoring within a task when additional practice is needed

Analysis and Research Purposes

Statistical programs with an intelligent advisor

Intelligent scheduling and calibrating of experimental items

Intelligent data collection for studies in school settings

Our imaginations will produce many promising applications, far more than can actually be developed. The development of such programs is a time- consuming and costly process. Acquiring the knowledge bases alone, from human experts, is very time consuming and resource intensive. The state of the art in expert systems is not far enough advanced to assure success in each undertaking.

Despite the difficulties of the undertaking, certain IM applications will indeed be developed. This chapter has narrowed its focus to the automation of test-administration processes, and three of the more promising IM applications are discussed.

THREE POTENTIAL CONTRIBUTIONS OF IM TO TEST ADMINISTRATION

Of the three promising contributions of IM to test administration discussed in this section, the first two are more likely to be of practical use in the near future. The third is more complex, but it represents a natural progression from the third generation.

Intelligent measurement can use machine intelligence to (a) score complex constructed responses involved in items and in reference tasks, (b) generate interpretations based on individual profiles of scores, (c) provide prescriptive advice to learners and teachers, to optimize progress through a curriculum. These contributions will be discussed in order.

Intelligent Scoring of Complex Constructed Responses

A knowledge base for scoring standards and rules can be used, along with automatic inferencing procedures, to provide the basis for automating complex scoring processes that now require costly human time. There is a natural pressure in educational measurement to move beyond decontextualized multiple-choice items toward other, more contextualized, item types. Two sorts of pressure always exist to broaden the types of items in the psychometric arsenal. The major form of pressure, from the scientific point of view, is for improved construct validity. The constructs involved in expertise relevant in real-world settings are what we seek: the roots of valued human performance relevant to social roles. We can do this best by modeling more accurately the critical aspects of work situations standardized for measurement, in which the constructs involved in expertise are required for success. Measurement organizations are criticized for reducing complex domains of human expertise to the knowledge aspects that can be tested with items requiring only a selection from alternatives. Multiple-choice items sample knowledge domains efficiently, but, as the cognitive scientists point out, declarative knowledge (knowing what) and procedural knowledge (knowing how) are two very different things. Without adding to the argument of how much procedural knowledge can be assessed by limited-choice items, moving to reference tasks in the third generation offers considerable promise. These tasks refer to actual performances in valued human roles. They offer greater potential for requiring procedural knowledge, along with corresponding increases in construct validity. The use of reference tasks might also reduce the second kind of pressure, that from a concerned user public who sees greater face validity in joblike or lifelike reference tasks than in decontextualized knowledge items. The distinction between competence and mastery in the third generation is relevant to this discussion. Both measurement scientists and the user public want measurement to reflect behavior samples that possess greater face and construct validity, as related to valued human mastery. Minimum competence is not enough. Neither is the sampling of factual knowledge adequate. Complex constructed responses more closely resemble what masters do.

In moving beyond limited-choice knowledge items to complex constructed response items, finding the scoring models for each constructed response item or reference task presents a problem for educational measurement. Such tasks have the stimulus standardized and loosen the standardization of the response. Each scoring model must assign values in a meaningful way to important variations in the complex response. The values must be assigned in a way that adequately models increments in expertise in the construct or constructs measured by the task. It is also useful to retain additional information to help identify intermediate or erroneous cognitive structures evidenced by the examinee's performance. This information can be used to guide prompting and coaching in CM applications or to guide interpretation and counseling.

Complex scoring models are routinely developed by testing organizations, but they are not presently replicated in automated systems. Testing organizations bring human experts together to spend many hours discussing how to score each of several constructed response items holistically and how to assign incremental points. For example, in the College Board Advanced Placement programs and in some of the Graduate Record Examination subject tests, ETS provides disciplined holistic scoring of constructed response items in the form of written essays or written protocols describing the solution of problems requiring mathematics. These problems partake of some of the attributes of mastery: individual and unique productions. The items in some of these examinations meet several parts of the definition of reference tasks given earlier. These items and their scoring models are standardized. They become reliable through the application of established, disciplined, holistic scoring methods. Applied artificial intelligence, through the new tool of expert systems, might substantially reduce the labor required to read tens of thousands of items with constructed responses. (Note, however, that it does not reduce the intelligent labor required to reach agreement on the scoring model.) In so doing, applied artificial intelligence might offer some of the benefits of mass scoring of multiple-choice items. Scoring could be used for both institutional and individual purposes, because the system could be programmed to provide feedback and repeated practice to learners on similar items and to produce a reliable and construct-valid score for institutional uses.

A project conducted by Bennett, Gong, Kershaw, and Rock (1988) examined this possibility in the context of the advanced placement program for computer science. Students currently deliver a program written in the Pascal language. They may submit it on a floppy disk, because a text file, rather than handwriting on a piece of paper, is the normal mode for editing a program on a computer. Bennett, Gong, Kershaw, and Rock are working with Elliot Soloway of Yale University to examine the applicability of his artificially intelligent program, called PROUST (Johnson & Soloway, 1985), for grading these questions automatically. The results indicate that the scoring can indeed occur automatically. It will be necessary to develop human quality-control procedures over the whole process and to develop a "manufacturing technology" to routinely capture the knowledge of experts and the variations in student behavior in standard Pascal programming situations.

As discussed in connection with the first generation, human observation and judgment in assigning the holistic score are necessary whenever vocalizations, skilled movements, unique written productions, or unique artistic productions are required. These responses inhere in much of what is valuable in the world, life, and work. By utilizing expert systems to score these responses, educational measurement can move beyond competence toward mastery for larger numbers of individuals in the future.

Automation of Individual Profile Interpretations

Human counselors and other professionals routinely examine profiles of scores and provide interpretive commentary for individuals to aid in career and vocational counseling, diagnosis of learning strengths and weaknesses, and placement decisions. Many of these experts have built up a base of experience and knowledge that can be captured through techniques of knowledge acquisition and programmed as an intelligent advice giver that mimics their expertise. The input would be the profile of scores, perhaps available in the same computer system that administered the tests through a CT and CAT battery. The output could include a series of questions that the counselor might ask to clarify ambiguous points. An interpretive commentary might then be printed out as a small set of the most likely pieces of advice. The professional could edit this initial draft if needed.

Intelligent Advice During Continuous Measurement

Providing intelligent advice during learning is the most promising contribution of IM for students and teachers. Its goal is the optimization of learning. It requires a curriculum administered in association with a continuous measurement delivery system. It requires that human expertise be acquired in a computerized knowledge base, analogous to that of the expert counselors who interpret individual profiles of static scores. The difference is that, in CM, the measurement is dynamic. This makes the knowledge more complex, but the validation easier. The knowledge is complex because of the many variations of individual trajectories and individual learner profiles. The validation is easier because the measurement is continuous, and the results of decisions at one level are immediately known at the next level.

Providing intelligent advice during continuous measurement is the epitome of computerized educational measurement. The optimization of learning in a growth space of calibrated educational tasks represents a challenge for educational measurement scientists and practitioners and it will require great effort over many years.

INTELLIGENT TUTORS: A CONVERGING OR DISCONTINUOUS LINE OF DEVELOPMENT?

Intelligent tutoring is a current application of machine intelligence that does not fit into either a familiar educational or measurement framework. Sleeman and Brown (1982) and Kearsley (1987) provide a variety of examples of intelligent tutors applied in different subject matter areas. Intelligent tutors are a relatively recent development that does not yet intersect with measurement thinking. The two fields have not been related to one another and have not benefited much from the work of one another. It is doubtful that the two fields will come together, unless, like Embretson, Tatsuoka, Sternberg and Glaser, measurement scientists encompass cognitive components and strategies in their models. Vaguely defined constructs inferred from aptitude factors have not proven useful to intelligent tutor developers. Even John Frederiksen, who has strong psychometric credentials, has not yet found a psychometric approach of value to his work with intelligent tutors (Frederiksen & White, in press). Rather, the electronic troubleshooting tutor these authors have constructed tutors learners from module to module, motivating each module with the earlier ones and building a cognitive foundation for each new module in the process. No measurement scales are needed. Few developers of intelligent tutors, however, have come to grips with how to deal with individual differences in anything but a discrete and categorical way, or with how to use measurement to more fully validate their constructs and claims.

Intelligent tutors are contributing important insights, research vehicles, and models for both the third and fourth generations. Some intelligent tutors of narrow scope are modules that could be treated as single lessons. These would be of interest in a continuous measurement system for dealing with the coaching, practice, and feedback inherent within a single reference task module. Working with these specific modules, measurement scientists and intelligent tutor developers could jointly define experimental and field- testing conditions for obtaining data relevant to the validation of the cognitive constructs in an intelligent tutoring model. The feedback, coaching, and repeated practice mechanisms in these models are examples of advanced tools for instruction. These tools could be used within the framework of a mastery map in a continuous measurement system, thus putting them in the framework of an entire course.

A few intelligent tutors have been implemented as entire courses. These systems are actually prototype continuous measurement systems with fourth-generation attributes. Unlike the examples of TICCIT and the WICAT Reading Curricula discussed earlier, they do not emphasize sequence control and status information made visible through the mastery map of the domain. They are less concerned with flexibility in sequencing through the domain, but simply establish a structured curriculum consisting of an ordered series of tasks that culminate in course mastery.

Two such courses have been in operation for over ten years at the Institute of Mathematical Studies in the Social Sciences at Stanford University. The oldest is a computer-aided education course in axiomatic set theory (Suppes & Sheehan, 1981a). The youngest, a CAE course in logic, is also in use at other universities (Suppes & Sheehan, 1981b). Both of these intelligent tutoring courses use automatic theorem proving for checking the correctness of students' proofs in axiomatic set theory or symbolic logic. The feedback the computer gives in the course of checking students' proofs enables the computer-student dialogue to continue and the students to learn from their errors. Students also have access to graduate student proctors in the machine room. An ambitious new project to apply these methods to precollege calculus is currently under way under NSF funding (Suppes et. al. 1987).

Another multiyear effort that has resulted in entire courses using artificial intelligence and intelligent tutoring is being conducted under the direction of John Anderson at Carnegie-Mellon University. The LISP tutor (Anderson, in press) is an extremely interesting intelligent tutor based on Anderson's ACT* Theory of Cognition (Anderson, 1983). This theory makes claims about the organization and acquisition of complex cognitive skills. The LISP tutor currently teaches a full-semester, self-paced course. It both tests the claims of ACT* theory and, simultaneously, provides university students with automated instruction in LISP programming. While Anderson has found that students working on problems with the LISP tutor get a letter grade higher on final exams than students not working with it, he does not claim that students do as well as they would with a human tutor.

The LISP tutor uses a mechanism called model tracing. A model is one of hundreds of ideal and buggy production rules. A production is an if-then statement: IF the goal is to x, THEN use the LISP function y and set the subgoal to z. Buggy rules have this same form. By tracing what the student is doing and matching it to one of these correct or buggy rules, the LISP tutor is able to generate helpful feedback messages. These feedback messages enable a student using a buggy rule to get back on the right track. Within both the same and subsequent lessons, the student might have many opportunities to practice on a given production. Data collected from the LISP tutor show an initial and dramatic drop in learning a given production. This validates the ACT* Theory, which predicts that the knowledge is initially "compiled." Learning after the knowledge is compiled seems to fit the standard power law of practice.

In analyzing the data from the LISP tutor, Anderson (in press) attempted to trace the source of individual differences in learning productions. He found two major factors, one dealing with acquisition and the other with retention.

The opportunities for measurement in these complete courses involving intelligent tutors are extensive, as they are in other computerized educational systems. It is a significant challenge, however, to analyze all of the data that can be generated. As researchers interested in psychometrically modeling the dynamic changes in learning become involved with such systems, we can hope for a convergence of different scientific approaches, rather than a totally diverging line of development.

COMPLICATIONS OF ARTIFICIAL INTELLIGENCE: FUTURE GENERATIONS

Artificial intelligence can deliver so much control and initiative into the hands of the user that the conditions for standardized measurement become impossible to achieve. This challenge of artificial intelligence is inherent in the concept of an intelligent curriculum. It is possible to implement in a computer system another kind of expert knowledge base, that of the domain expert or experts, those who now write the textbooks and teach the classes constituting the curriculum. The term intelligent curriculum means that the curriculum must include such an expert knowledge base. Students will have access to an inference engine that can answer queries based on the expert knowledge in the domain. Students will be able to perform searches through and query the knowledge base, and the system will be able to answer these queries in a manner approximating that of human experts. Students will be able, at some point, to add the results of these queries to the knowledge base and build a richer personalized system for answering queries from an individual line of investigation. This scenario goes beyond intelligent tutoring projects, which use carefully structured tasks, to a more substantial manipulation of entire textbooks and other sources of educational content.

This kind of creative, fluid behavior in investigating and manipulating knowledge in a new way has some of the properties of mastery defined earlier. Students in this role will be manipulating and adding to knowledge in a personal way. This is an exciting prospect for education, but it will complicate the possible contributions of educational measurement. Scales of growing competence might be difficult or impossible to develop in an intelligent curriculum. There are no standardized items, testlets, or reference tasks. Where there is no standardization, there is no measurement. The tasks and abilities currently familiar to the educational measurement community might cease to be of much interest to the community at large, and new tasks and abilities could be extremely hard to measure.

One avenue of approach would be to consider the work of searching, querying, and adding to the knowledge base as a very complex reference task in and of itself. The construct measured would then be the new types of learning and problem-solving expertise required to use computerized knowledge bases of subject matter domains. Standardized tasks could be developed that require use of the knowledge base to produce different, personally constructed, productions. Disciplined subjective scoring protocols could then be developed for each standardized task. This approach might take us to the (perhaps more tractable) concepts of IM use for intelligent scoring of constructed responses.

SUMMARY

This chapter was written at a time of dynamic change in the field of computer-administered measurement. It deals with some trends that are apparent and in prospect. So that this chapter will not become quickly dated, the strategy used was to describe four generations of computerized educational measurement, all based on rapidly emerging technological tools. As a result of the wide availability and low cost of new technological delivery systems, test delivery is shifting from paper-and-pencil and printed booklets to on-line computer workstations. The technologies that make this possible include:

1. Low cost and high computing and storage capabilities of newly available technologies.
2. Hardware and software to provide the communication between workstations or response stations and a single computer in which records can be kept for everyone in a group.
3. Availability of large-capacity optical memories, such as videodiscs and CD-ROMS, which allow the wide distribution of curriculum and testing materials of great scope at low cost. This development also permits the mediation of testing and teaching presentations by means of video, audio, computer graphics, and text.
4. Development of networking capabilities to distribute testing displays and collect responses in a central location.
5. Developments in psychometric procedures for calibrating tasks and estimating the position of individuals on scales (item

response theory and required extensions).

6. Developments in knowledge-based computing and expert systems for building and querying interactive knowledge bases.

Together, these technologies expand and permit partial replication of the human capabilities of sensing, remembering, deciding, acting, and communicating. Before computer administration, these processes were implemented through mark-sense sheets and scanners, computer scoring and reporting of scanned test sheets, manual administration and scoring of individually administered tests, and disciplined holistic scoring of constructed responses. The four generations permit automation of these processes in new ways, with greater potential efficiency.

The first generation of computer-aided testing enables us to do what we now do, but to do it faster, more accurately, and with much more interesting and realistic displays and responses. What we do now, in the main, is to take static measurements for institutional purposes. Computerized testing and CAT enhance these purposes and make possible some additional applications for individuals.

The second generation provides a new theory and adds considerable efficiency to the administration of computerized tests. The calibration of items makes possible the adaptive selection of items during test administration. Adaptive presentation, based on dynamic adjustment of the display or response time or adaptive arrangement of content, is also possible.

The third generation, continuous measurement, offers potential discontinuity from current methods in the practice of educational measurement, educational research, and teaching. The distinction between testing and curriculum begins to fade. Measurement becomes unobtrusive. Development of educational measurements will combine with curriculum development, and educational research will combine with educational practice. The CM generation offers learners and teachers continuous monitoring of progress on mastery maps of the domain to be mastered and the finer grained monitoring of progress within reference tasks, so that advice can be given to aid the teacher and the system in providing coaching to guide further practice. Individuals' trajectories through the domain represented in the mastery map will be available. Learner profiles will emerge made of both generic measures and new measures of approach, avoidance, and strategy within the system.

The third generation will not fully achieve its goals without the new tools provided by the knowledge bases and inference engines supplied by the fourth generation. Intelligent measurement will make possible adaptive and intelligent advice based on individual trajectories and learner profiles. Before this goal is achieved, machine intelligence will be used to score complex constructed responses automatically and to provide complex interpretations of individual profiles made of static measurements.

Future generations might include fully intelligent curricula: knowledge bases which can be queried and expanded by users skilled in new learning and discovery strategies for using such symbolically represented expertise. In this case, the role of educational measurement will be reduced or will shift to the measurement of new forms of expertise in learning, problem solving, and individual expansion and reorganization of the knowledge domain.

CONCLUDING THOUGHTS

GENERATIONAL ENHANCEMENTS IN POWERS OF OBSERVATION

The technological developments of these four generations confer upon the educational community the possibility of increased powers of measurement and, thus, increased powers of observation. These increased powers make visible the previously invisible. This yields better information and specification, which, in turn, leads to expansion of the field of inquiry.

The significance of using new technologies to enhance powers of observation can be put in historical perspective by recalling the introduction of, for example, the microscope and x-ray technology. When van Leeuwenhoek, a lens grinder, looked through his microscope at his sperm and his spit and saw "cavorting beasties" for the first time, his powers of observation were enhanced by a newly discovered lens grinding technology. Whole new fields of science, technology, and human service have evolved from this technology and its refinement. The biological classification of life was revised to add fungi, protista, and monera to the plant and animal kingdoms. As optical microscopes were improved and the electron microscope developed, the fields of microbiology and genetics and new material sciences evolved.

The second example of technology is the X-ray. Here powers of observation of the interior of living organisms and other opaque objects were dramatically enhanced. A host of specialties within medical and dental science use diagnostic methods based on this enhanced observational power.

In both of these cases, technological innovation made visible the previously invisible. Powers of observation were enhanced and magnitudes of the newly observed phenomena were scaled and measured.

Will technological enhancements of powers of observation lead to similar breakthroughs in educational theory and practice? Our belief that they will is closely tied to a particular view of the ends of educational research and the role of measurement in fostering those ends: Educational research is the study of the trajectories of growth over time. Its major goal is the identification of key attributes that govern growth and improvement and prevent decay and deterioration. Measurement, on the other hand, is the quantitative specification of the position, direction, and velocity of an individual or group of individuals in an educationally relevant growth space. The goals of educational research are therefore dependent on measurement. The practice of measurement is significantly advanced by the introduction of computerized test administration, because, with the judicious application of hardware, software, and psychometric technologies, the specification of position and velocity in growth space can be accomplished.

REFERENCES

The topics discussed in this chapter are a part of a rapidly moving field. To have restricted the references to refereed journals or books would have narrowed the scope of the chapter to an unduly conservative position. Reports and other references not found in the public literature can be obtained in most cases from the authors.

Abernathy, L. J. (1986). Computerized placement tests: A revolution in testing instruments. New York: College Board.

Ager, T. A. (in press). From interactive instruction to interactive testing. In R. Freedle (Ed.). Artificial intelligence and the future of testing. Hillsdale, NJ: Lawrence Erlbaum.

Alderman, D. L. (1978). Evaluation of the TICCIT computer-assisted instruction system in the community college. Princeton, NJ: Educational Testing Service.

American Psychological Association, Committee on Professional Standards (COPS) and Committee of Psychological Tests and Assessment (CPTA) (1986). Guidelines for computer based tests and interpretations. Washington, DC: Author.

Anderson, J. R. (1983) The architecture of cognition. Cambridge, MA: Harvard University Press.

Anderson, J. R. (in press). Analysis of student performance with the LISP tutor. In N. Frederiksen, R. Glaser, A. Lesgold, & M. Shafto (Eds.), Diagnostic monitoring of skill and knowledge acquisition. Hillsdale, NJ: Lawrence Erlbaum.

Andrich, D. (1985). A latent trait model for items with response dependencies: Implications for test construction and analysis. In S. E. Embretson (Ed.), Test design: Developments in psychology and psychometrics. New York: Academic Press.

Assessment Systems Corporation. (1987). User's manual for the MicroCAT Testing System. St. Paul: Author.

Baker, F. B. (in press). Computer technology in test construction and processing. In R. L. Linn (Ed.), Educational Measurement (3rd ed.). New York: Macmillan.

Baker, F. B. (1984). Technology and testing: State of the art and trends for the future. Journal of Educational Measurement, 21, 399-406.

Bennett, R.E., Gong, B., Kershaw, R.C., Rock, D.A., Soloway, E., & Macalalad, A. (1988). Agreement between expert system and human ratings of constructed responses to computer science items (Report No. RR-88-20). Princeton, NJ: Educational Testing Service.

Binet, A. (1909). Les idees modernes sur les enfants [Modern ideas about children]. Paris: Ernest Flamarion.

Birnbaum, A. (1968). Some latent trait models and their uses in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

Biskin, B. H. & Kolotkin, R. L. (1977). Effects of computerized administration on scores on the Minnesota Multiphasic Personality Inventory. Applied Psychological Measurement, 1(4), 543-549.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. Psychometrika, 46, 443-459.

Bock, R. D. & Mislevy, R. J. (1982a) Adaptive EAP estimation of ability in a micro computer environment. Applied Psychological Measurement, 6(4) 431-444.

Bock, R.D. & Mislevy, R.J. (1982b) BILOG: Maximum likelihood item analysis and test scoring with logistic ogive models. Mooresville, IN: Scientific Software.

Bock, R. D. & Mislevy, R. J. (1982c). Biweight estimates of latent ability. Educational and Psychological Measurement, 42, 725-737.

Bridgeman, B., Bennett R., & Swinton, S. (1986). Design of an interactive assessment videodisc demonstration project. Princeton, NJ: Educational Testing Service.

Bunderson, C. V. (1973). The TICCIT project: Design strategy for educational innovation. In S. A. Harrison & L. M. Stolurow (Eds.). Productivity in higher education. Washington, DC: National Institutes of Education.

Bunderson, C. V., & Faust, G. W. (1976). Programmed and computer assisted instruction, Chapter III. In N. L. Gage (Ed.). Seventy-Fifth Yearbook, The psychology of teaching methods (pp. 44-90). Chicago, IL: The National Society for the Study of Education.

Bunderson, C. V., & Inouye, D. K. (1987). Computer-aided educational delivery systems. In R. Gagne (Ed.), Instructional technology. Hillsdale, NJ: Lawrence Erlbaum.

Calvert, E. J., & Waterfall, R. C. (1982). A comparison of conventional and automated administration of Raven's Standard Progressive Matrices. International Journal of Man-Machine Studies, 17, 305-310.

Cassirer, E. (1923). Substance and function and Einstein's theory of relativity (W. C. Swabey & M. C. Swabey, Trans.). New York: Dover.

Dillon, R. F., & Stevenson-Hicks, R. (1981). Effects of item difficulty and method of test administration on eye span patterns during analogical reasoning (Technical report No. 1, Contract N66001-80-C0467). Carbondale, IL: Southern Illinois University.

Druesne, B., Kershaw, R., & Toru, O. (1986). CLEP general examination: Humanities. Princeton, NJ: Educational Testing Service.

Eignor, D. R. & Cook, L. L. (1983). An investigation of the feasibility of using item response theory in the preequating of aptitude tests. Paper presented at the meeting of the American Educational Research Association, Montreal.

Elwood, D. L. (1972a). Automated WAIS testing correlated with face-to-face testing: A validity study. International Journal of Man-Machine Studies, 4, 129-137.

Elwood, D. L. (1972b). Test retest reliability and cost analysis of automated and face-to-face intelligence testing. International Journal of Man-Machine Studies, 4, 1-22.

Embretson, S. (1983). Psychometrics for theory based tests. Paper presented at the meeting of the American Educational Research Association, Montreal.

Embretson, S. E. (1985a). Multicomponent models for test design. In S. E. Embretson (Ed.). Test design: Developments in psychology and psychometrics. New York: Academic Press.

Embretson, S. E. (1985b). (Ed.) Test design: Developments in psychology and psychometrics. New York: Academic Press.

Embretson, S. (in press). Diagnostic testing by measuring learning processes: Psychometric considerations for dynamic testing. In N. Fredericksen, R. Glaser, A. Lesgold, M. Shafto (Eds.). Diagnostic monitoring of skill and knowledge acquisition. Hillsdale, NJ: Lawrence Erlbaum.

Feurzeig, W. & Jones, G. (1970). Reevaluating low achievers with computer-administered tests. Unpublished manuscript, Bolt, Beranek, & Newman, Boston.

Forehand, G. A. & Bunderson, C. V. (1987a). Basic concepts of mastery assessment systems. Princeton, NJ: Educational Testing Service.

Forehand, G. A. & Bunderson, C. V. (1987b). Mastery assessment systems and educational objectives. Princeton, NJ: Educational Testing Service.

Freedle, R. (Ed.). (in press). Artificial intelligence and the future of testing. Hillsdale, NJ: Lawrence Erlbaum.

Frederiksen, N., Glaser, R., Lesgold, A., & Shafto, M. (Eds.) (in press). Diagnostic monitoring of skill and knowledge acquisition. Hillsdale, NJ: Lawrence Erlbaum.

Frederiksen, J., & White, F. (in press). Intelligent tutors as intelligent testers. In N. Frederiksen, R. Glaser, A. Lesgold, & M. Shafto (Eds.). Diagnostic monitoring of skill and knowledge acquisition. Hillsdale, NJ: Lawrence Erlbaum.

Gagne, R. M. (1965). The conditions of learning, 1st ed. New York: Holt, Rinehart & Winston.

Gagne, R. M. (1968). Learning Hierarchies. Educational Psychologist, 6, 1-9.

Gardner, D. P. (1983). A nation at risk: The imperative for educational reform. Washington, DC: National Commission on Excellence in Education.

Glaser, R. (1986). The integration of instruction and testing. The Redesign of Testing for the 21st Century: Procedings of the 1985 ETS Invitational Conference. Princeton, NJ: Educational Testing Service.

Greaud, V. A. & Green, B. F. (1986). Equivalence of conventional and computer presentation of speed tests. Applied Psychological Measurement, 10(1), 23-24.

Green, B. F. (1983). The promise of tailored tests. In H. Wainer & S. Messick (Eds.), Principals of modern psychological measurement. Hillsdale, NJ: Lawrence Erlbaum.

Green, B. F. (1988). Construct validity of computer-based tests. In H. Wainer & H. Braun (Eds.), Test validity. Hillsdale, NJ: Lawrence Erlbaum.

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., & Reckase, M. D. (1982). Evaluation plan for the computerized adaptive vocational aptitude battery (Research Report No. 82-1). Baltimore: Johns Hopkins University.

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., & Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive tests. Journal of Educational Measurement, 21, 347-360.

Green, B. F., Bock, R. D., Linn, R. L., Lord, F. M., & Reckase, M. D. (1983). A plan for scaling the computerized adaptive ASVAB. Baltimore: Johns Hopkins University.

Hambleton, R. K. (in press) Principles and selected applications of item response theory. In R. L. Linn (Ed.), Educational Measurement (3rd ed.). New York: Macmillan.

Hambleton, R. K., & Swaminathan, H. (1985). Item response theory. Boston: Kluwer-Nijhoff Publishing.

Hedl, J. J., O'Neil, H. F., & Hansen, D. N. (1971, February). Computer based intelligence testing. Paper presented at the meeting of the American Educational Research Association, New York.

Heuston, D. (1985). Some considerations affecting the use of computers in public education. Provo, UT: WICAT Systems.

Hitti, F. J., Riffer, R. L., & Stuckless, E. R. (1971). Computer-managed testing: A feasibility study with deaf students. Rochester, NY: Rochester Institute of Technology, National Technical Institute for the Deaf.

Hoffman, K. I. & Lundberg, G. D. (1976). A comparison of computer-monitored group tests with paper-and-pencil tests. Educational and Psychological Measurement, 36, 791-809.

Holtzman, W. H. (Ed.), (1970). Computer assisted instruction, testing and guidance. New York: Harper & Row.

Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). Item response theory: Application to psychological measurement. Homewood, IL: Dow Jones-Irwin.

Hunt, E. (1986). Cognitive research and future test design. The Redesign of Testing for the 21st Century: Proceedings of the 1985 ETS Invitational Conference. Princeton, NJ: Educational Testing Service.

Inouye, D. K., & Bunderson, C. V. (1986). Four generations of computerized test administration. Machine-Mediated Learning, 1, 355-371.

Inouye, D. K., & Sorenson, M. R. (1985). Profiles of dyslexia: The computer as an instrument of vision. In D. B. Gray & J. F. Kavanagh (Eds.). Biobehavioral measures of dyslexia. Parkton, MD: York Press.

Jacobs, R. L., Byrd, D. M., & High, W. R. (1985). Computerized testing: The hidden figures test. Journal of Educational Computing Research, 1(2), 173-177.

Johnson, D. F., & Mihal, W. L. (1973). Performance of blacks and whites in computerized versus manual testing environments. *American Psychologist*, *28*, 694-699.

Johnson, W. L. & Soloway, E. (1985). PROUST: An automatic debugger for Pascal programs. *BYTE*, *10*(4), 179-190.

Jonassen, D. H. (1986, January). *Effects of microcomputer display on a perceptual/cognitive task*. Paper presented at the meeting of the Association for Educational Communications and Technology, Las Vegas, NV.

Jones, D. H. (1982). *Redescending M-type estimators of latent ability* (Research Tech. Rep. No. 82-30). Princeton, NJ: Educational Testing Service.

Kearsley, G. P. (Ed.), (1987). *Artificial intelligence & instruction*. Reading, MA: Addison-Wesley.

Kiely, G. L., Zara, A. R., & Weiss, D. J. (1986). *Equivalence of computer and paper-and-pencil Armed Services Vocational Aptitude Battery Tests*. (Research Report No. AFHRL-TP-86-13). Brooks Air Force Base, TX: Air Force Human Resources Laboratory.

Kingston, N. M. & Dorans, N. J. (1984). Item location effects and their implications for IRT equating and adaptive testing. *Applied Psychological Measurement*, 146-154.

Kirsch, I. S., & Jungeblut, A. (1986). *Literacy: Profiles of America's young adults--Final Report* (NAEP Report No. 16-PL-01). Princeton, NJ: National Assessment of Educational Progress.

Knights, R. M., Richardson, D. H., & McNarry, L. R. (1973). Automated vs. clinical administration of the Peabody Picture Vocabulary Test and the Coloured Progressive Matrices. *American Journal of Mental Deficiency*, *78*(2), 223-225.

Koch, W. R. & Reckase, M. D. (1978). *A live tailored testing comparison study of the one- and three-parameter logistic models* (Research Report 78-1). Columbia, MO: University of Missouri, Tailored Testing Research Laboratory, Educational Psychology Department.

Koson, D., Kitchen, C., Kochen, M., & Stodolosky, D. (1970). Psychological testing by computer: Effect of response bias. *Educational and Psychological Measurement*, *30*, 803-810.

Lee, J. A. & Hopkins, L. (1985, April). *The effects of training on computerized aptitude test performance and anxiety*. Paper presented at the meeting of the Eastern Psychological Association, Boston.

Lee, J., Moreno, K. E., & Sympson, J. B. (1986). The effects of mode of test administration on test performance. Educational and Psychological Measurement, 46, 467-474.

Linacre, M., & Wright, B. D. (1984). MICROSCALE. Chicago: MESA Press.

Linn, R. L. (1986). Barriers to new test designs. The Redesign of Testing for the 21st century: Proceedings of the 1985 ETS Invitational Conference. Princeton, NJ: Educational Testing Service.

Lord, F. M. (1952). A theory of test scores. Psychometric Monographs, No. 7.

Lord, F. M. (1970). Some test theory for tailored testing. In W. H. Holtzman (Ed.), Computer-assisted instruction, testing and guidance. New York: Harper & Row.

Lord, F. M. (1977). Practical applications of item characteristic curve theory. Journal of Educational Measurement, 14, 117-138.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum.

Lord, F. M., & Novick, M. R. (1968). Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

Luce, R. D. (1986). Response times, their role in inferring elementary mental organization. New York: Oxford University Press.

Lukin, M. E., Dowd, T., Plake, B. S., & Kraft, R. G. (1985). Comparing computerized versus traditional psychological assessment. Computers in Human Behavior, 1, 49-58.

Lushene, R., O'Neil, H., & Dunn, T. (1974). Equivalent validity of a completely computerized MMPI. Journal of Personality Assessment, 13, 407-412.

Mazzeo, J., & Harvey, A. L. (1988). The equivalence of scores from automated & conventional versions of educational & psychological tests: A review of the literature (Research Report No. CBR-87-8, ETS RR-88-21). Princeton, NJ: Educational Testing Service.

McBride, J. R., & Moe, K. C. (1986, April). Computerized adaptive achievement testing. Paper presented at the meeting of the National Council for Measurement in Education, San Francisco.

McBride, J. R., & Weiss, D. J. (1974). A word knowledge item pool for adaptive ability measurement (Research Report No. 74-2). Minneapolis: Minnesota University.

McKinley, R. L., & Reckase, M. D. (1980). Computer applications to ability testing. Association for Educational Data Systems Journal, 13, 193-203.

McKinley, R. L., & Reckase, M. D. (1984a, April). Implementing an adaptive testing program in an instructional program environment. Paper presented at the meeting of the American Educational Research Association, New Orleans.

McKinley, R. L., & Reckase, M. D. (1984b). A latent trait model for use with sequentially arranged units of instruction (Research Report ONR 84-2). Iowa City, IA: American College Testing Program.

Merrill, M. D. (1983). Component display theory. In C. M. Reigeluth (Ed.), Instructional design theories and models: An overview of their current status, Hillsdale, NJ: Lawrence Erlbaum.

Merrill, M. D. Schneider, E. W., and Fletcher, K. A. (1980). TICCIT. Englewood Cliffs, NJ: Educational Technology Publications.

Millman, J. (1977, April). Creating domain referenced tests by computer. Paper presented at the meeting of the American Educational Research Association, New York.

Millman, J. (1980). Computer-based item generation. In R. Berk (Ed.) Criterion referenced measurement: The state of the art. Baltimore: Johns Hopkins University Press.

Millman, J. (1984a, April). Computer-assisted test construction. Paper presented at the meeting of the American Educational Research Association, New Orleans.

Millman, J. (1984b). Individualizing the construction and administration of tests by computer. In R. Berk (Ed.), A guide to criterion-referenced test construction. Baltimore: Johns Hopkins University Press.

Milson, R., Lewis, M., & Anderson, J. R. (in press). The teacher's apprentice project: Building an algebra tutor. In R. Freedle (Ed.) Artificial intelligence and the future of testing, Hillsdale, NJ: Lawrence Erlbaum.

Mislevy, R. J., & Verhelst, N. (1987). Modeling item responses when different subjects employ different solution strategies (Research Report No. 87-47-ONR). Princeton, NJ: Educational Testing Service.

Mislevy, R. J., & Bock, R. D. (1982). Maximum likelihood item analysis and test scoring with binary logistic models. Mooresville, IN: Scientific Software.

Moreno, K., Wetzel, C. D., McBride, J. R., & Weiss, D. J. (1983). Relationship between corresponding Armed Services Vocational Aptitude Battery (ASVAB) and computerized adaptive testing (CAT) subtests (NPRDC TR 83-27). San Diego: Navy Personnel Research and Development Center.

National Board of Medical Examiners. (1987). Update on computer based testing. National Board Examiner, 34, 1-3.

Olsen, J. B., & Bunderson, C. V. (1980). Toward the development of a learner profile battery: Theory and research. Orem, UT: WICAT Systems.

Olsen, J. B., Bunderson, C. V., & Gilstrap, R. M. (1982). Development of a preliminary design for a computerized adaptive testing system for the Department of Defense. Orem, UT: WICAT Systems.

Olsen, J. B., Bunderson, C. V., & Gilstrap, R. M. (1983). Prototype design and development of a computerized adaptive testing system. Orem, UT: WICAT Systems.

Olsen, J. B., Inouye, D., Hansen, E. G., Slawson, D. A., & Maynes, D. M. (1984). The development and pilot testing of a comprehensive assessment system. Provo, UT: WICAT Education Institute.

Olsen, J. B., Maynes, D. M., Ho, K., & Slawson, D. A. (1986). The development and pilot testing of a comprehensive assessment system, phase I. Provo, UT: Waterford Testing Center.

Olsen, J. B., Maynes, D. M., Slawson, D. A., & Ho, K. (1986, April). Comparison and equating of paper-administered, computer-administered and computerized adaptive tests of achievement. Paper presented at the meeting of the American Educational Research Association, San Francisco.

Owen, R. J. (1969). A Bayesian approach to tailored testing (Research Bulletin No. 69-92). Princeton, NJ: Educational Testing Service.

Owen, R. J. (1975). A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. Journal of the American Statistical Association, 70, 351-356.

Parks, B. T., Mead, D. E., & Johnson, B. L. (1985). Validation of a computer administered marital adjustment test. Journal of Marital and Family Therapy, 11(2), 207-210.

Patience, W. M., & Reckase, M. D. (1979). Operational characteristics of a one-parameter tailored testing procedure (Research Report No. 79-2). Columbia, MO: University of Missouri, Tailored Testing Research Laboratory, Educational Psychology Department.

Psychological Corporation. (1986). Computerized adaptive differential aptitude test. San Antonio, TX: Author.

Quan, B. L., Park, T. A., Sandahl, G., & Wolfe, J. H. (1984). Microcomputer network for computerized adaptive testing research (NPRDC TR 84-33). San Diego: Navy Personnel Research and Development Center.

Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. Copenhagen: Nielsen and Lydiche.

Reckase, M. D. (1985). Models for multidimensional and hierarchically structured training materials (Research Report No. ONR 85-1). Iowa City, IA: American College Testing Program.

Reckase, M. D., Carlson, J. E., & Ackerman, T. A. (1986, August). The effect of computer presentation on the difficulty of test items. Paper presented at the meeting of the American Psychological Association, Washington, DC.

Reigeluth, C. M., Merrill, M. D., & Bunderson, C. V. (1978). The structure of subject matter content and its instructional design implications. Instructional Science, 7, 107-126.

Robertson, J. R., Inouye, D. K., & Olsen, J. B. (1985). Basic skills testing system. Provo, UT: Waterford Testing Center.

Rock, D., & Pollack, J. (1987). Measuring gains: A new look at an old problem. Princeton, NJ: Educational Testing Service.

Sachar, J. D., & Fletcher, J. D. (1978). Administering paper-and-pencil tests by computer, or the medium is not always the message. In D. J. Weiss (Ed.), Proceedings of the 1977 Computerized Adaptive Testing Conference. Wayzata, MN: University of Minnesota.

Samejima, F. (1977). A comment on Birnbaum's three-parameter logistic model in the latent trait theory. Psychometrika, 38, 221-233.

Scissons, E. H. (1976). Computer administration of the California Psychological Inventory. Measurement and Evaluation in Guidance, 9(1), 22-25.

Shephard, R. N., & Metzler, J. (1971). Mental rotation of three-dimensional objects. Science 171, 701-703.

Slawson, D. A. (1986). District wide computerized assessment in Texas. Provo, UT: Waterford Testing Center.

Slawson, D. A., Maynes, D. M., Olsen, J. B., & Foster, D. F. (1986). Waterford test creation package. Provo, UT: Waterford Testing Center.

Sleeman, D., & Brown, J. S. (1982). Intelligent tutoring systems. New York: Academic Press.

Sternberg, R. J. (1977). Intelligence, information processing, and anological reasoning: The componential analysis of human abilities. Hillsdale, NJ: Lawrence Erlbaum.

Sternberg, R. J. (1982). Reasoning, problem solving and intelligence. In R. J. Sternberg (Ed.), Handbook of human intelligence. Cambridge: Cambridge University Press.

Sternberg. R. J., & McNamara, S. M. (1985). The representation and processing of information in real-time verbal comprehension. In S. E. Embretson, (Ed.), Test design: Developments in psychology and psychometrics. New York: Academic Press.

Suppes, P., Ager, T., Berg P., Chuaqui, R., Graham, W., Maas, R. E., & Takahashi, S. (1987). Applications of computer technology to pre-college calculus (Technical Report No. 310, Psychology in Education Series). Stanford, CA: Stanford University, Institute for Mathematical Studies in the Social Sciences.

Suppes, P., & Sheehan, J. (1981a). CAI course in axiomatic set theory. In P. Suppes (Ed.), University-level computer assisted instruction at Stanford: 1968-1980 (pp.3-80). Stanford, CA: Stanford University, Institute for Mathematical Studies in the Social Sciences.

Suppes, P., & Sheehan, J. (1981b). CAI course in logic. In P. Suppes (Ed.), University-level computer assisted instruction at Stanford: 1968-1980 (pp.193-226). Stanford, CA: Stanford University, Institute for Mathematical Studies in the Social Sciences.

Tatsuoka, K. (in press). Toward an integration of item response theory and cognitive error diagnosis. In N. Fredericksen, R. Glaser, A. Lesgold, & M. Shafto, (Eds.). Diagnostic monitoring of skill and knowledge acquisition. Hillsdale, NJ: Lawrence Erlbaum.

Thissen, D. (1986). MULTILOG [Computer program]. Mooresville, IN: Scientific Software.

Thissen, D., & Steinberg, L. (1984). A response model for multiple-choice item. Psychometrika, 49, 501-519.

Vale, C. D., & Gialluca, K. A. (1985). ASCAL: A microcomputer program for estimating logistic IRT item parameters. St. Paul: Assessment Systems Corporation.

Wainer, H. (1983). On item response theory and computerized adaptive tests. Journal of College Admissions, 27, 9-16.

Wainer, H. (1984). The development of computerized adaptive testing system (D-CATS). Unpublished manuscript, Educational Testing Service, Princeton, NJ.

Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. Journal of Educational Measurement, 24(3), pp. 195-201.

Wainer, H., & Thissen, D. M. (1987). Estimating ability with the wrong model. Journal of Educational Statistics, 12 (4), pp. 339-368.

Wainer, H., & Wright, B. D. (1980). Robust estimation of ability in the Rasch model. Psychometrika, 45, 373-391.

Ward, W. C. (1984). Using microcomputers to administer tests. Educational Measurement: Issues and Practice, 3, 16-20.

Ward, W. C. (1986). Measurement research that will change test design for the future. The Redesign of Testing for the 21st Century: Proceedings of the 1985 ETS Invitational Conference. Princeton, NJ: Educational Testing Service.

Ward, W. C., Kline, R. G., & Flaugher, J. (1986). College Board Computerized Placement Tests: Validation of an adaptive test of basic skills (Report No. PR-86-29). Princeton, NJ: Educational Testing Service.

Watts, K., Baddeley, A., & Williams, M. (1982). Automated tailored testing using Raven's matrices and the Mill Hill vocabulary tests: A comparison with manual administration. International Journal of Man-Machine Studies, 17, 331-344.

Weiss, D. J. (1985). Computerized adaptive measurement of achievement and ability (Final report N00014-79-C-0172). Minneapolis: University of Minnesota, Computerized Adaptive Testing Laboratory.

Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. Journal of Educational Measurement, 21, 361-375.

White, D. M., Clements, C. B. & Fowler, R. D. (1985). A comparison of computer administration with standard administration of the MMPI. Computers in Human Behavior, 1, 153-162.

Whitely, S. E. & Dawis, R. V. (1976). The influence of test context on item difficulty. Educational and Psychological Measurement, 36, 329-337.

WICAT Education Institute. (1983). High technology and basic skills in reading. Provo, UT: Author.

WICAT Systems (1988). Learner profile and WICAT test of basic skills. Orem, UT: Author.

Wise, L. A., & Wise, S. L. (1986). Comparison of computer-administered and paper-administered achievement tests with elementary school children. Unpublished manuscript, University of Nebraska Lincoln, Department of Educational Psychology.

Wise, S. L., Boettcher, L. L., Harvey, A. L., & Plake, B. S. (1987, April). Computer-based testing versus paper-pencil testing: Effects of computer anxiety and computer experience. Paper presented at the meeting of the American Educational Research Association, Washington, DC.

Wise, S. L., Plake, B. S., Boettcher, L. L., Eastman, L. A., & Lukin, M. E. (1987, April). The effects of item feedback on test performance and anxiety in a computer-administered test. Paper presented at the meeting of the American Educational Research Association, Washington, DC.

Wingersky, M. S., Barton, M. A., & Lord, F. M. (1982). LOGIST User's Guide. Princeton, NJ: Educational Testing Service.

Wright, B. D., Rossner, M., & Congdon, R. (1984). M-SCALE. Chicago: MESA Press.

Yen, W. M. (1980). The extent, causes, and importance of context effects on item parameters for two latent trait models. Journal of Educational Measurement, 17, 297-311.

Yen, W. M. (1981). Using simulation results to choose a latent trait model. Applied Psychological Measurement, 5, 245-262.